Credit EDA Assignment

UMANG MALHOTRA

Problem Statement & Goal

Risks involved while Approving / Refusing a loan:

- Potential loss to the lending business, by refusing a loan to a client likely to repay, &
- Potential loss in a loan recovery, on approving a loan to client likely to default.

Goal to identify patterns and scenario where a loan is:

- ❑ Not rejected for an applicant capable of repaying the loan, and
- ❑ Not approved for an applicant having difficulties in repaying the loan.

TARGET variable:

Clients with payment difficulties and other cases.

Steps & Phases involved in EDA

Various Steps (not necessarily in sequence) undertaken for the assignment:

- Data Reading & Understanding
- Data Cleaning & Manipulation
- Data Analysis & Plotting Graphs
- Conclusions Drawn

Data Reading & Understanding

>>> Application & Previous dataset:

Observations:

- More than 3 and 16 lac records respectively
- Each record has 122 & 37 different features/variables
- A good mix of Numerical & Categorical features
- Missing values as 'XNA' & 'XAP' in a few categorical variables
- 'Unknown' and 'Other' categories observed in a few variables
- Considerable number of variables with null values
- Possible outliers observed: max CNT_CHILDREN (19), max AMT_APPLICATION (69L)
- Most common categories: Female, Married, and Cash loans

Data Cleaning & Manipulation

>>> Columns with significant null values:

Application Dataset:

- 49 columns with null values greater than 45%
- Most of these are about a client's building

Previous Application Dataset:

- 11 columns with null values greater than 40%
- Quite a few regarding days of loan drawing or due etc

Actions:

- ➢ Discarded columns with significant null values
- ➢ Merged both datasets
- ➢ Discarded quite a number of irrelevant columns

Data Cleaning & Manipulation

>>> Handle Missing Values:

Numerical Columns:

- Impute the missing values with the column mean, i.e. average value of the column
  - ❑ EXT_SOURCE_2/3,
  - ❑ AMT_REQ_CREDIT_BUREAU_YEAR/QRT/MON/WEEK/DAY/HOUR

- Impute the missing values with the column median i.e. the middle value among all
  - ❑ CNT_PAYMENT, AMT_ANNUITY_x/y, AMT_Credit_y

Categorical Columns:

- Impute the missing values by replacing with a new category 'Unknown'
  - ❑ OCCUPATION_TYPE

Data Cleaning & Manipulation

>>> Column Engineering:

Renaming columns to distinguish between data from present or previous application:

- AMT_ANNUITY, AMT_CREDIT, NAME_CONTRACT_TYPE

New columns derived:

Years of age derived from AGE specific columns, such as

- ✓ DAYS_BIRTH, DAYS_EMPLOYED

Average client rating can be derived from

- REGION_RATING_CLIENT & REGION_RATING_CLIENT_W_CITY

Data Cleaning & Manipulation

>>> Column Engineering:

Categorize columns to make appropriate value buckets/baskets:

- NAME_CASH_LOAN_PURPOSE, NAME_FAMILY_STATUS, NAME_HOUSING_TYPE, NAME_INCOME_TYPE, NAME_YIELD_GROUP, OCCUPATION_TYPE, ORGANIZATION_TYPE

Redundant columns dropped:

- REGION_RATING_CLIENT, REGION_RATING_CLIENT_W_CITY, NAME_CASH_LOAN_PURPOSE, DAYS_BIRTH, DAYS_EMPLOYED

Data Cleaning & Manipulation

>>> Impute Encoded missing values:

Replaced XNA/XAP values with column mode:

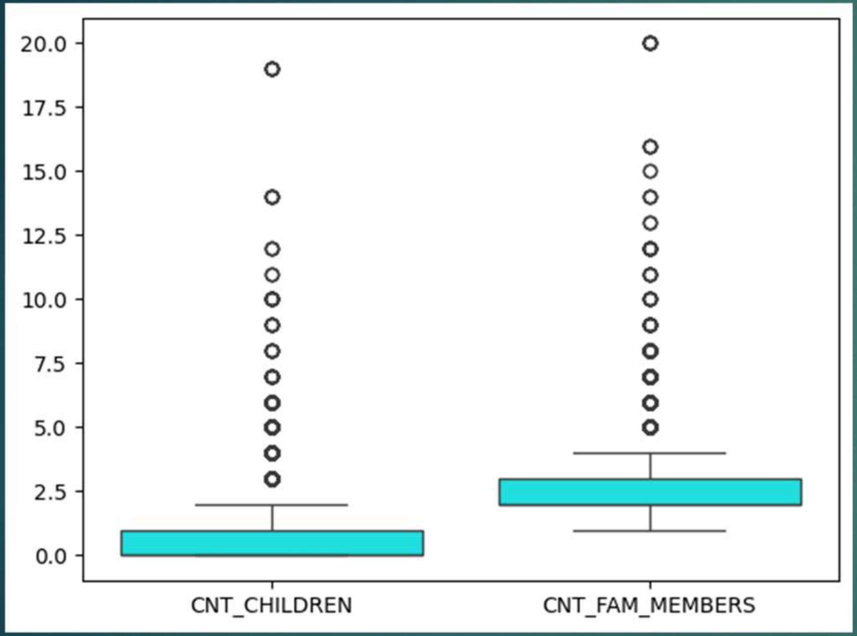
- CODE_GENDER, NAME_CONTRACT_TYPE_Prev,

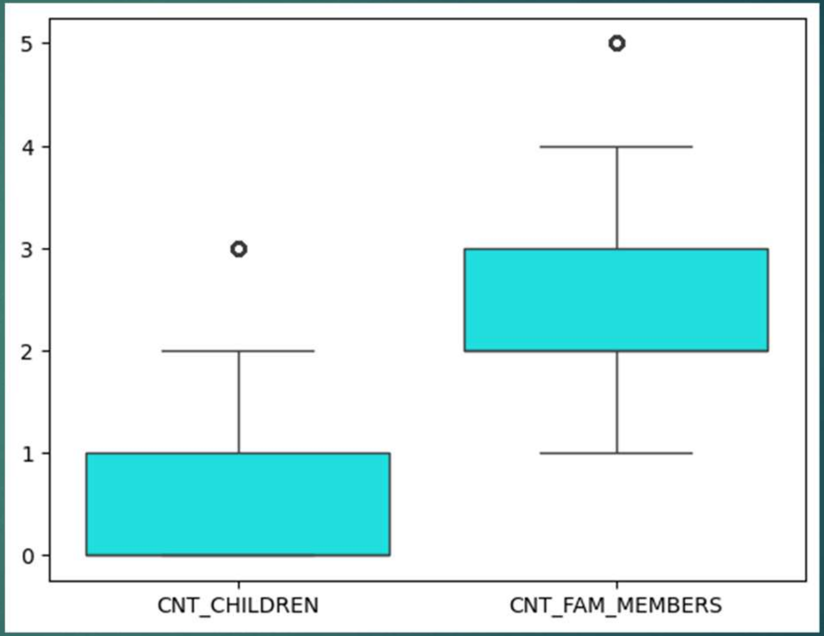
Impute the missing values by replacing with a new category 'Unknown':

- CODE_REJECT_REASON, NAME_YIELD_GROUP,

Data Cleaning & Manipulation

>>> Outliers:

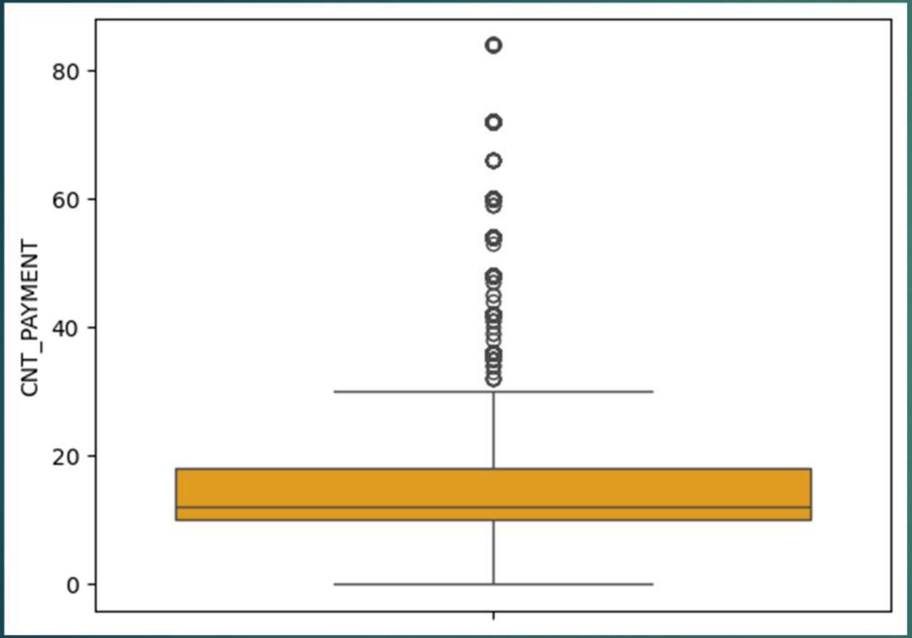


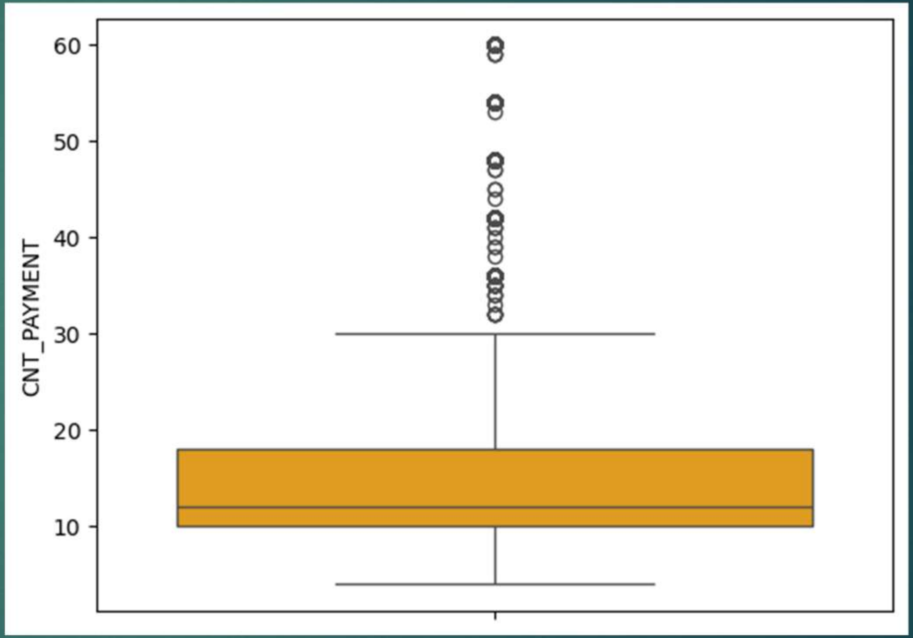


Up to 99th percentile, there are 3 or less children
We can limit the CNT_CHILDREN column up to 3

Originally, there were 5 or less family members, up to 99th percentile
Once the outliers in CNT_CHILDREN columns are discarded,
there are 5 or less family members at the max in CNT_FAM_MEMBERS
So, CNT_FAM_MEMBERS column is automatically limited up to 5

Data Cleaning & Manipulation >>> Outliers:

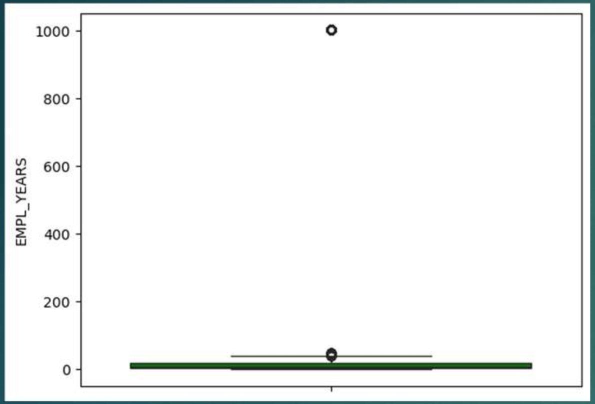


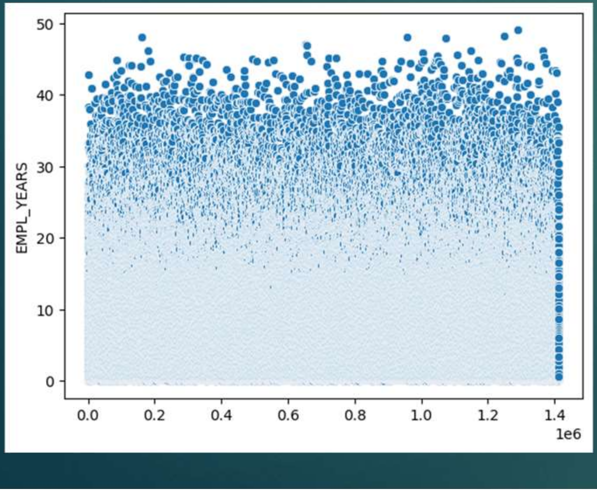


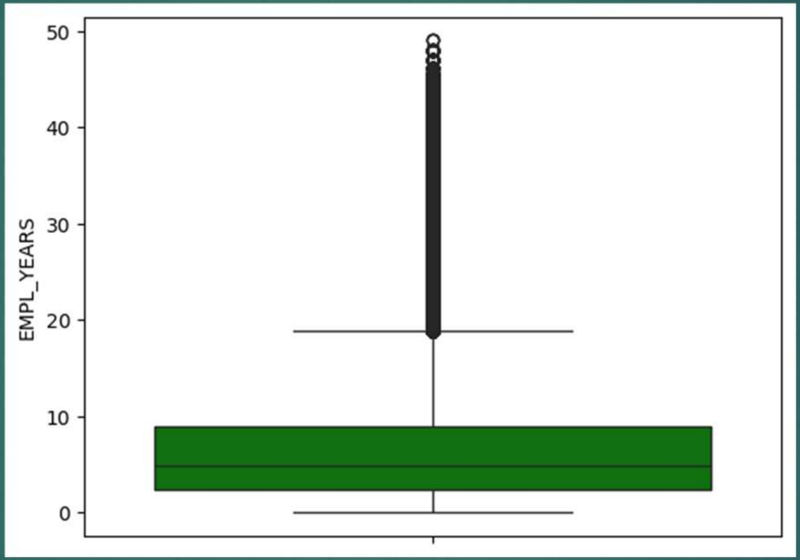
Lower bound below 0 is not practical value for a loan term.
Upper bound is quite below 99th percentile
CNT_PAYMENT values can be limited from 4 months to 60 months

Data Cleaning & Manipulation

>>> Outliers:



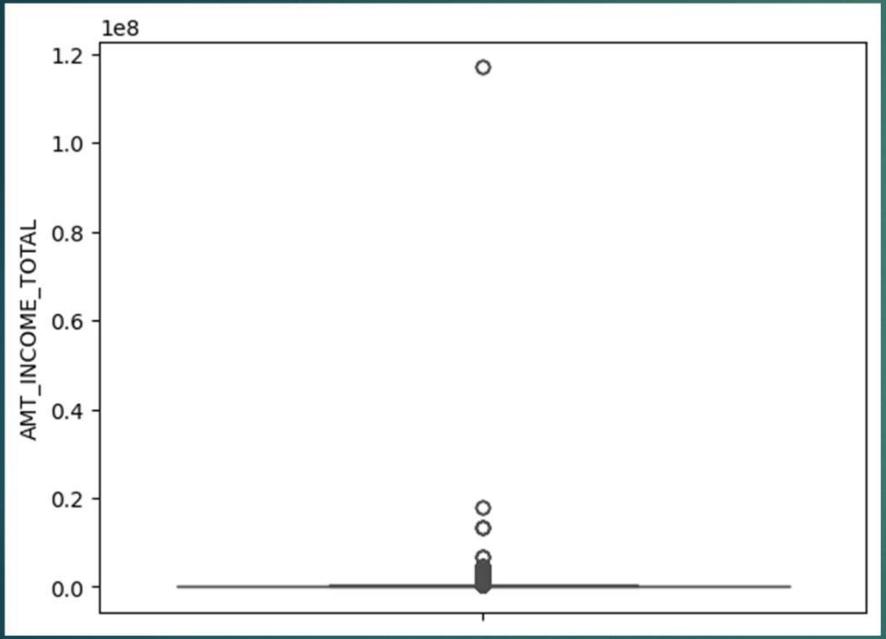


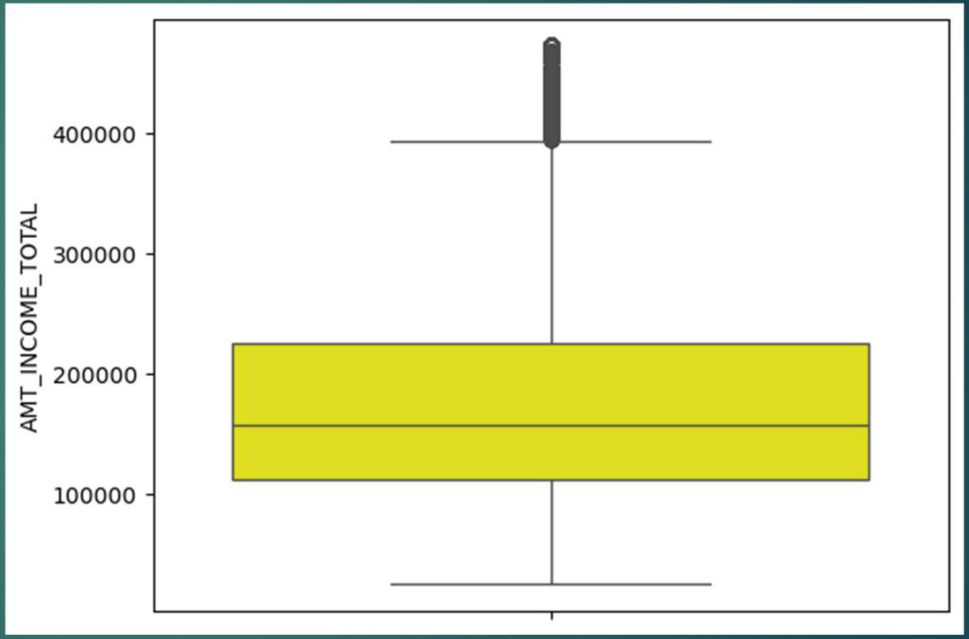


Most of the values lie within 50 in EMPL_YEARS column
The values can be limited within 50 for this column

Data Cleaning & Manipulation

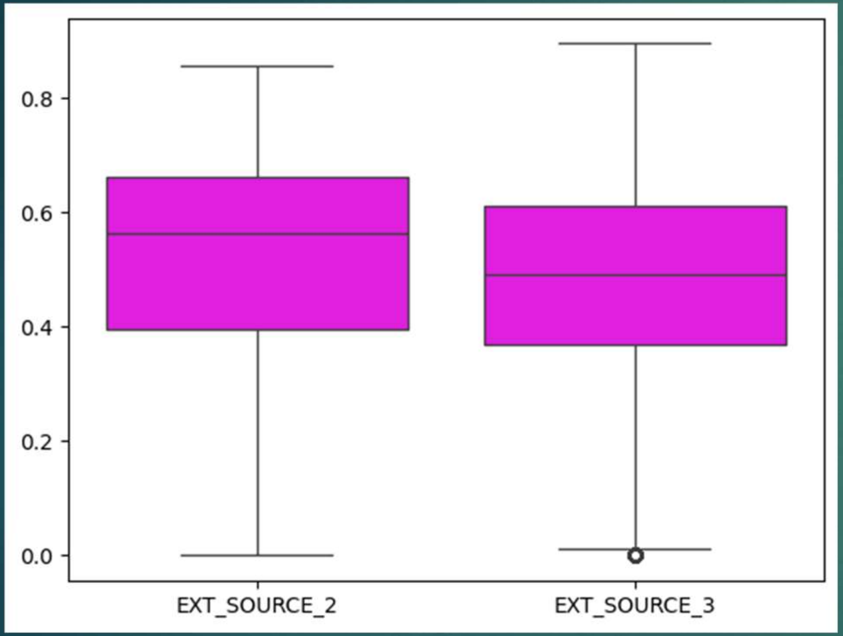
>>> Outliers:

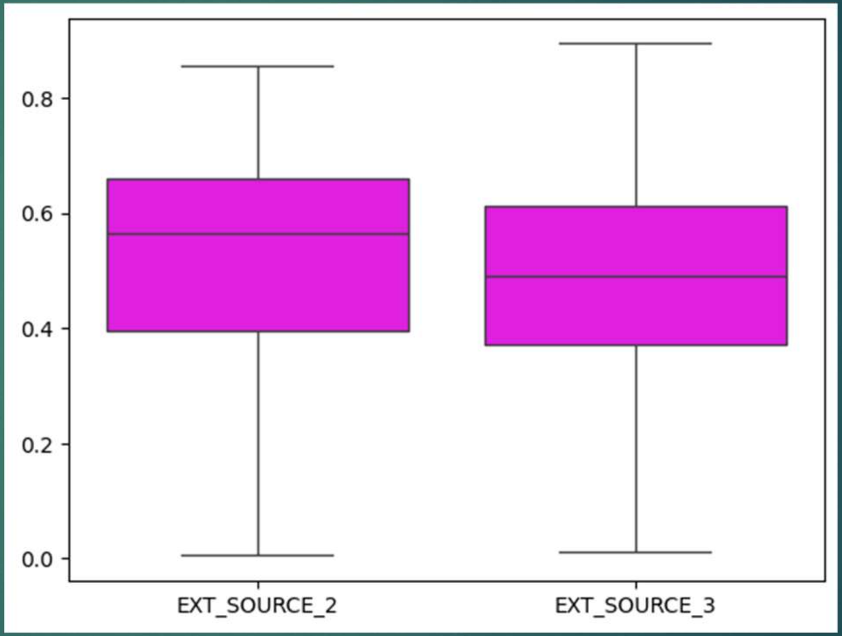




Up to 1% of the values in AMT_INCOME_TOTAL are beyond 99th percentile
The values in AMT_INCOME_TOTAL can be limited to 99th percentile

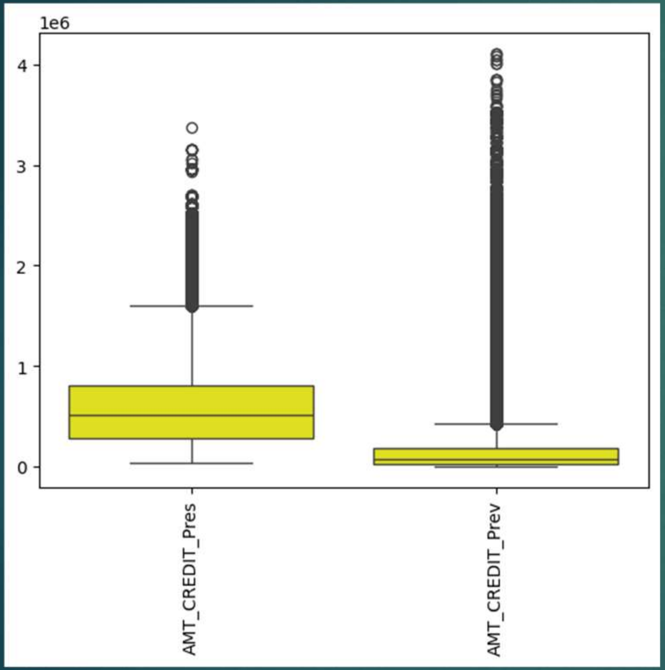
Data Cleaning & Manipulation >>> Outliers:

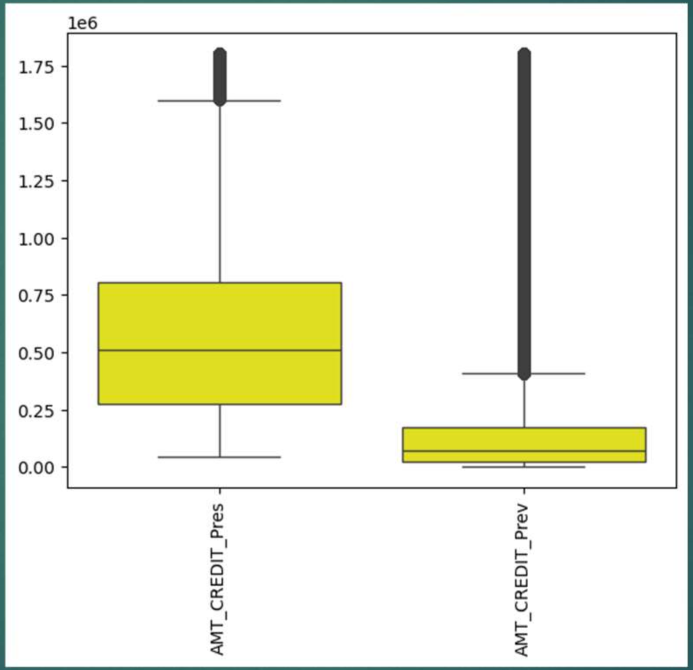




The minimum values in EXT_SOURCE_2 and EXT_SOURCE_2 columns are far less to be practical.
The external sources score values can be limited to around 0.005 in the lower side

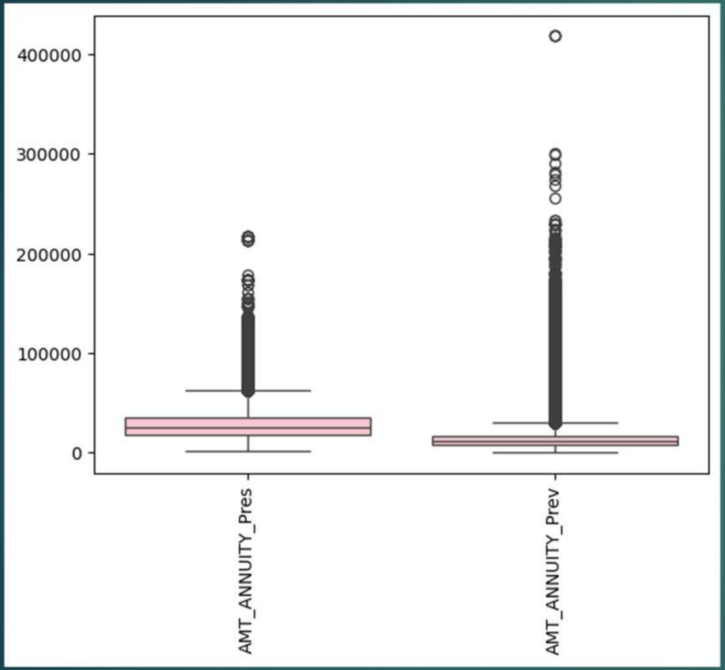
Data Cleaning & Manipulation >>> Outliers:

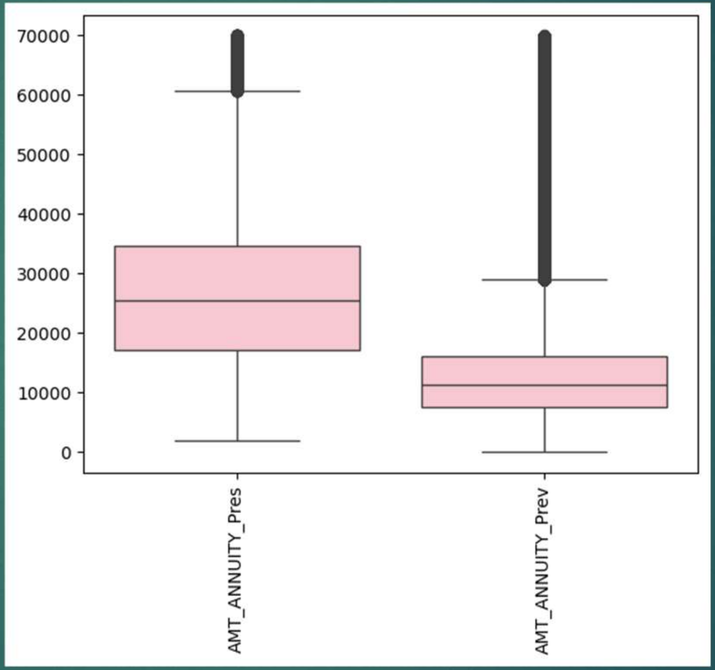




The CREDIT values can be limited up to 18L and 16L in present and previous cases respectively. This is quite close to their 90th percentile values.

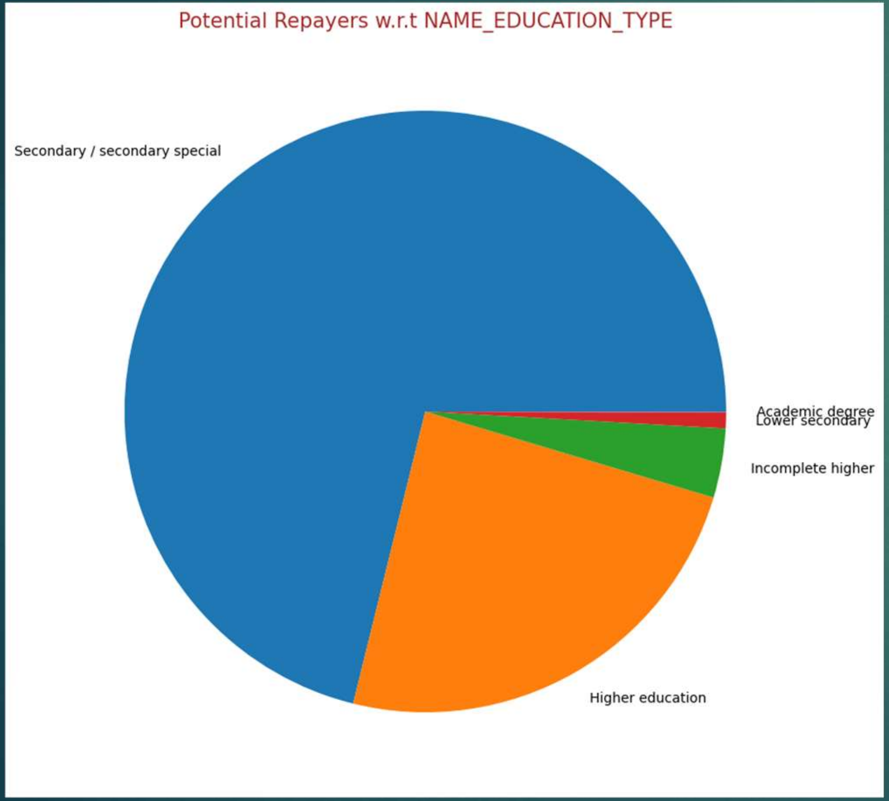
Data Cleaning & Manipulation >>> Outliers:





The ANNUITY values can be limited up to 70K in both present and previous cases. This is quite close to their 90th percentile values.

Data Analysis & Plotting Graphs
>>> Univariate Analyses - Ordinal:



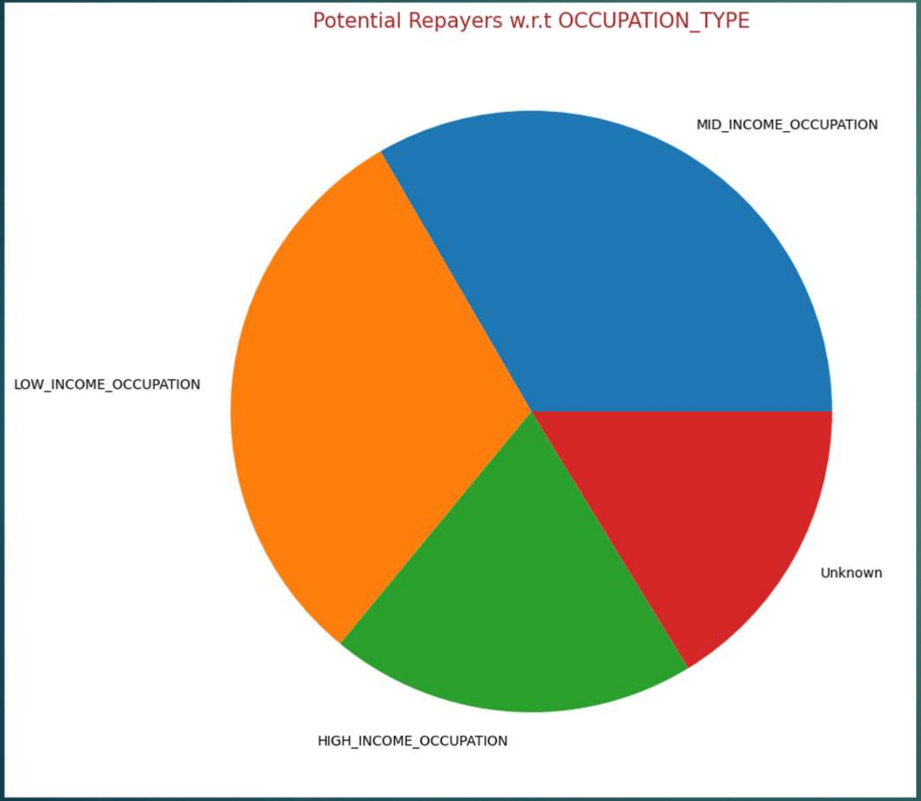
Applicants with an academic degree are the most probable of repaying their loans

Clients with secondary education are not very likely in their loan repayment

Disclaimer:
TARGET = 1 when a client has difficulty repaying loan
TARGET = 0 for other cases

Data Analysis & Plotting Graphs

>>> Univariate Analyses - Ordinal:



Clients in high income occupation such as IT/HR/Realty/Managers have more chances to repay a loan

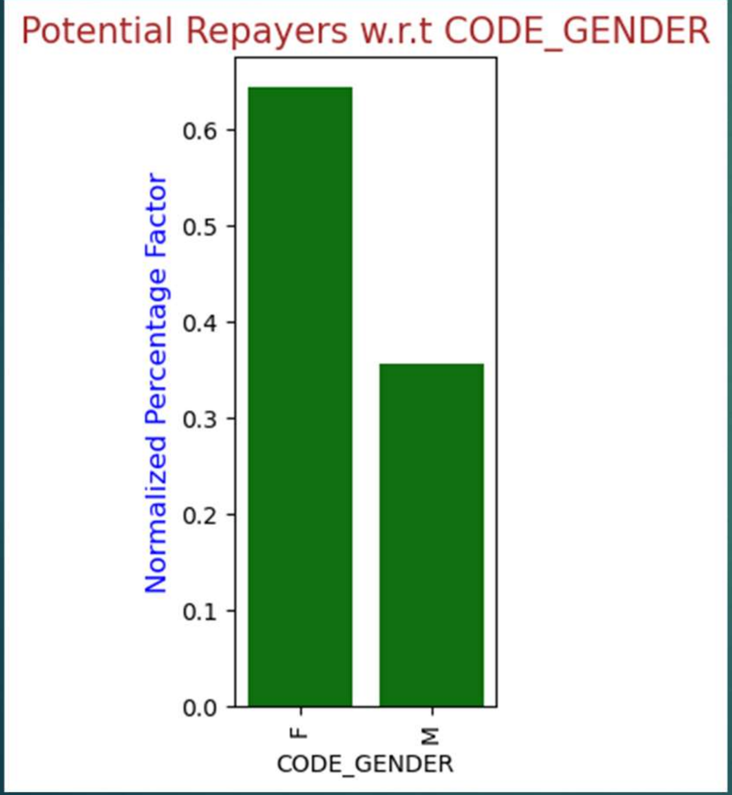
Applicants from low income group such as security/cooking/sales staff and drivers are more likely to default.

Disclaimer:
TARGET = 1 when a client has difficulty repaying loan
TARGET = 0 for other cases

Disclaimer: The insight is based on the available information, not considering ‘Unknown’ category as they could be in any of the occupation groups.

Data Analysis & Plotting Graphs

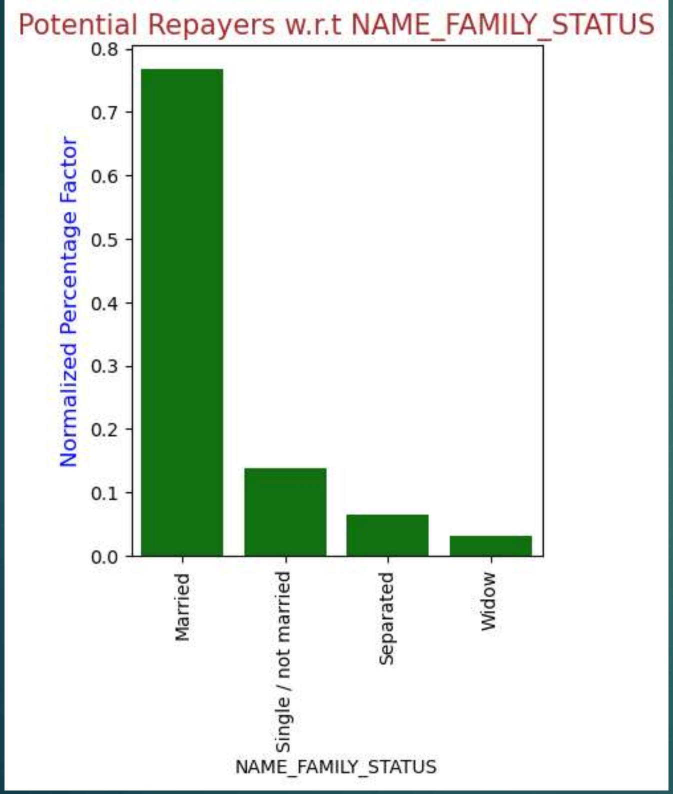
>>> Univariate Analyses - Nominal:



There is a higher probability of female clients repaying a loan as compared to male clients

Data Analysis & Plotting Graphs

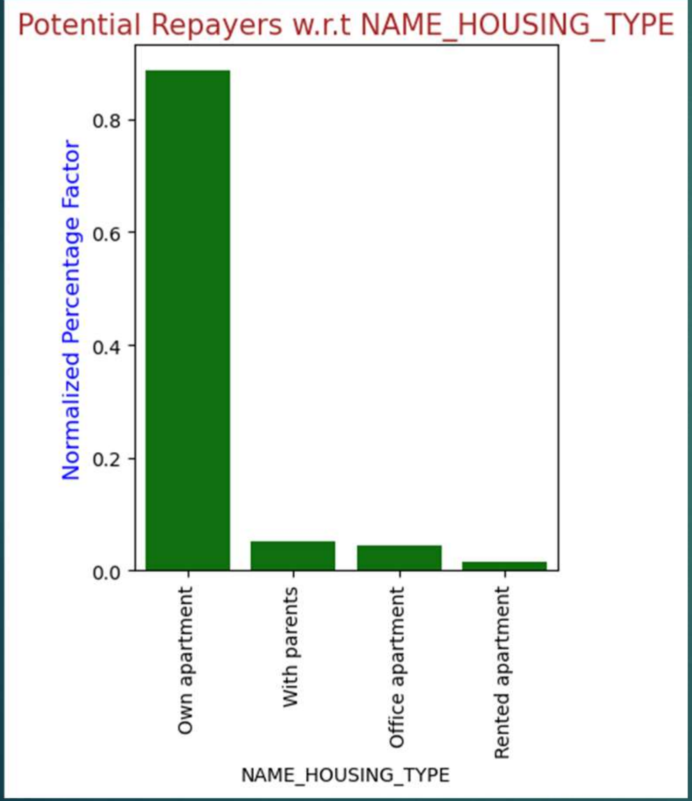
>>> Univariate Analyses - Nominal:



There is a higher probability of repaying a loan if a client is married as compared with the applicants who are single, separated, or a widow.

Data Analysis & Plotting Graphs

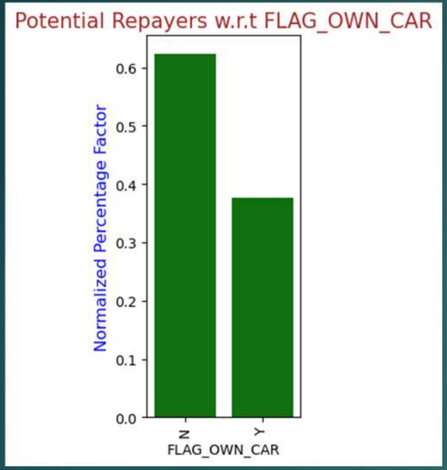
>>> Univariate Analyses - Nominal:

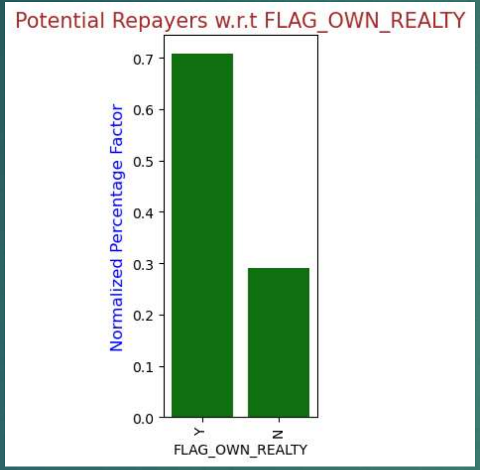


There is a higher probability of repaying a loan where a client has own apartment as compared to the applicants who are staying with parents, or in a rented or office apartment.

Data Analysis & Plotting Graphs

>>> Univariate Analyses - Nominal:



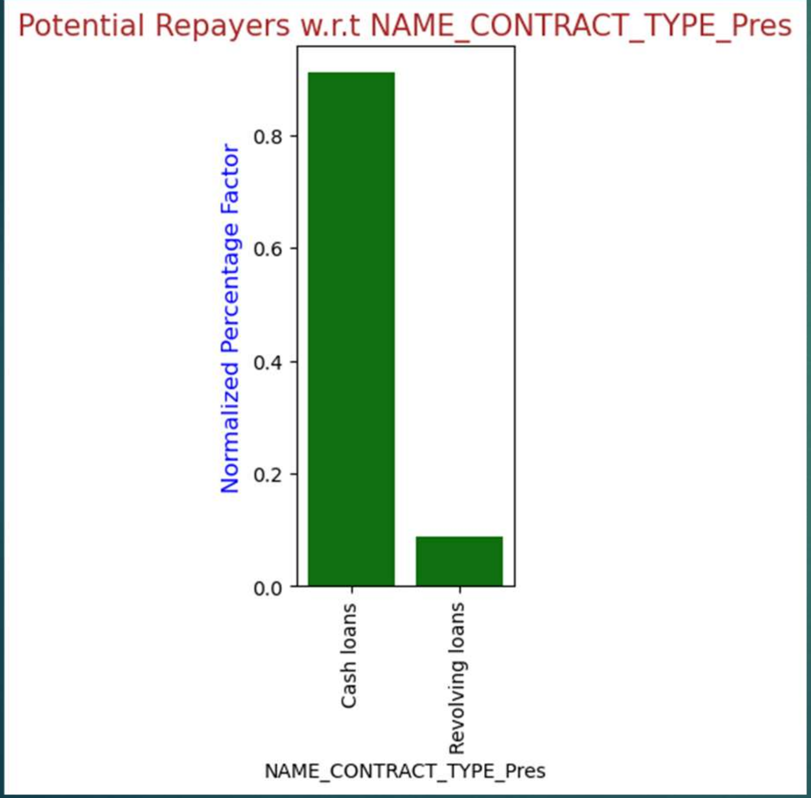


A client owning a property has more chances of repaying a loan.

The fact that a client has a car, doesn’t reflect good chances of their repaying a loan.

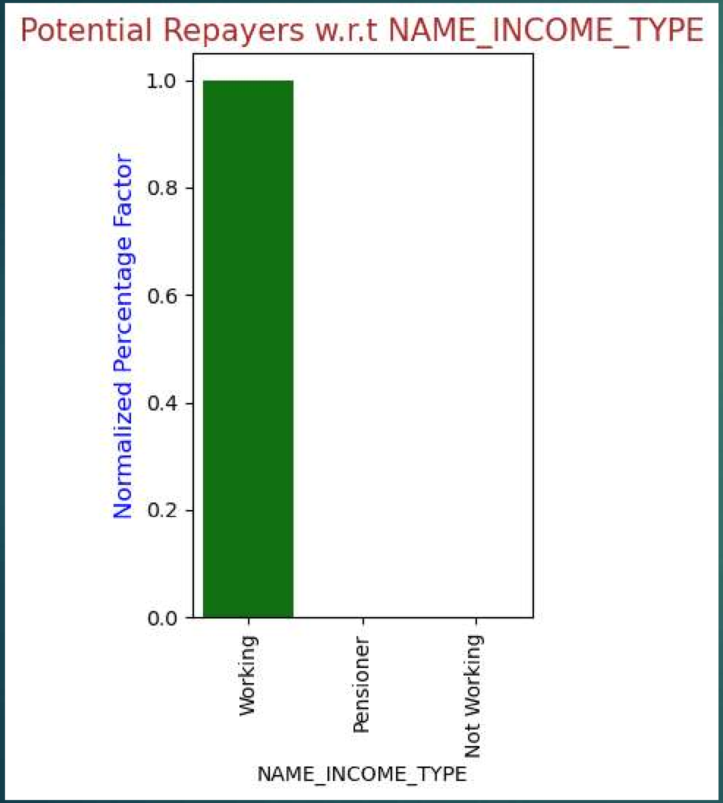
Data Analysis & Plotting Graphs

>>> Univariate Analyses - Nominal:



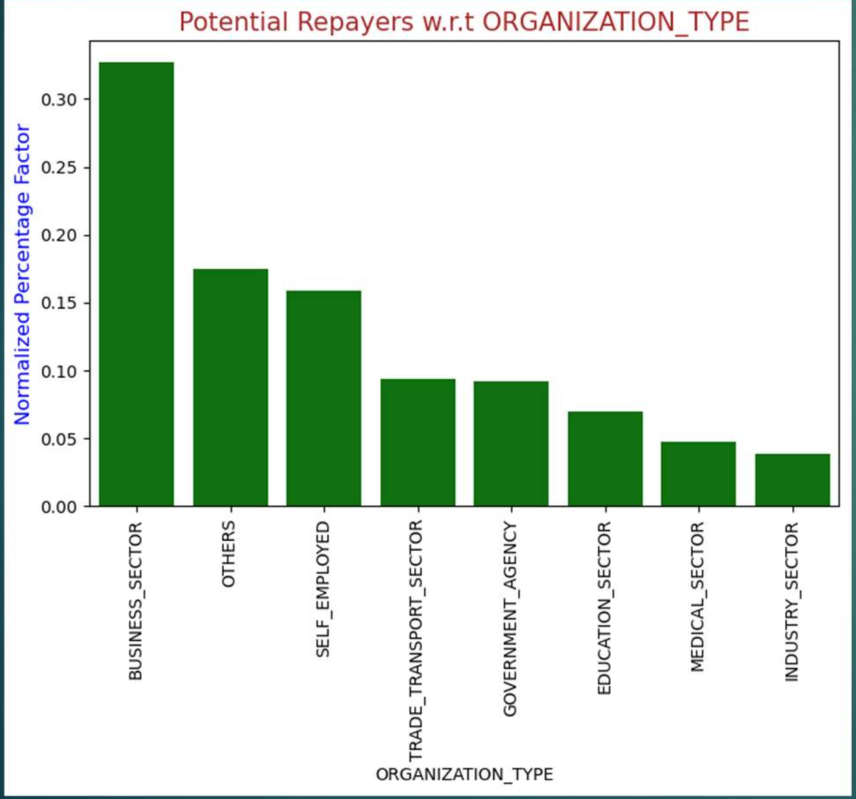
The clients who has applied for cash loans are more likely to repay their loans as compared to the applicants who applied for a revolving loan.

Data Analysis & Plotting Graphs
>>> Univariate Analyses - Nominal:



Quite naturally, the working applicants are going to have more probability of repaying their loans as compared to the clients who are not working such as pensioners.

Data Analysis & Plotting Graphs
>>> Univariate Analyses - Nominal:



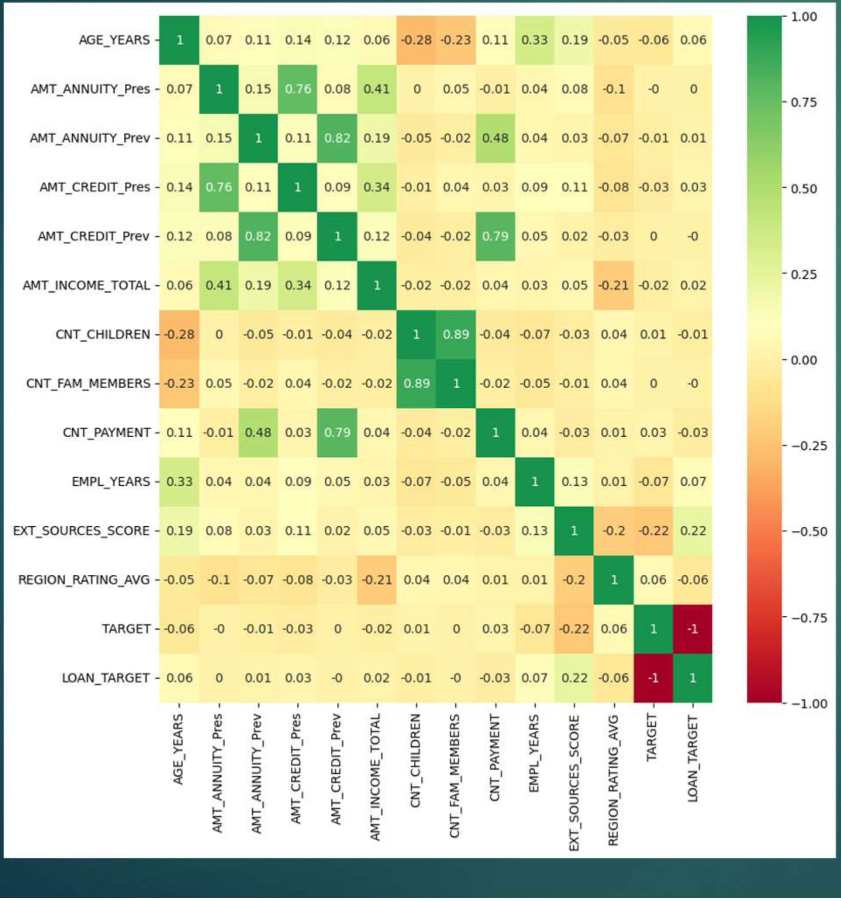
The clients in business sector are the most worthy of repaying their loans.

Even the applicants who are self employed have high chances of repaying their loans.

The clients who are working in trade, transport or industry sectors have quite low probability of repaying their loans.

Data Analysis & Plotting Graphs

>>> Bivariate Analyses - Numerical:



	AGE_YEARS	AMT_ANNUITY_Pres	AMT_ANNUITY_Prev	AMT_CREDIT_Pres	AMT_CREDIT_Prev	AMT_INCOME_TOTAL	CNT_CHILDREN	CNT_FAM_MEMBERS	CNT_PAYMENT	EMPL_YEARS	EXT_SOURCES_SCORE	REGION_RATING_AVG	TARGET	LOAN_TARGET
AGE_YEARS	1	0.07	0.11	0.14	0.12	0.06	-0.28	-0.23	0.11	0.33	0.19	-0.05	-0.06	0.06
AMT_ANNUITY_Pres	0.07	1	0.15	0.76	0.08	0.41	0	0.05	-0.01	0.04	0.08	-0.1	-0	0
AMT_ANNUITY_Prev	0.11	0.15	1	0.11	0.82	0.19	-0.05	-0.02	0.48	0.04	0.03	-0.07	-0.01	0.01
AMT_CREDIT_Pres	0.14	0.76	0.11	1	0.09	0.34	-0.01	0.04	0.03	0.09	0.11	-0.08	-0.03	0.03
AMT_CREDIT_Prev	0.12	0.08	0.82	0.09	1	0.12	-0.04	-0.02	0.79	0.05	0.02	-0.03	0	-0
AMT_INCOME_TOTAL	0.06	0.41	0.19	0.34	0.12	1	-0.02	-0.02	0.04	0.03	0.05	-0.21	-0.02	0.02
CNT_CHILDREN	-0.28	0	-0.05	-0.01	-0.04	-0.02	1	0.89	-0.04	-0.07	-0.03	0.04	0.01	-0.01
CNT_FAM_MEMBERS	-0.23	0.05	-0.02	0.04	-0.02	-0.02	0.89	1	-0.02	-0.05	-0.01	0.04	0	-0
CNT_PAYMENT	0.11	-0.01	0.48	0.03	0.79	0.04	-0.04	-0.02	1	0.04	-0.03	0.01	0.03	-0.03
EMPL_YEARS	0.33	0.04	0.04	0.09	0.05	0.03	-0.07	-0.05	0.04	1	0.13	0.01	-0.07	0.07
EXT_SOURCES_SCORE	0.19	0.08	0.03	0.11	0.02	0.05	-0.03	-0.01	-0.03	0.13	1	-0.2	-0.22	0.22
REGION_RATING_AVG	-0.05	-0.1	-0.07	-0.08	-0.03	-0.21	0.04	0.04	0.01	0.01	-0.2	1	0.06	-0.06
TARGET	-0.06	-0	-0.01	-0.03	0	-0.02	0.01	0	0.03	-0.07	-0.22	0.06	1	-1
LOAN_TARGET	0.06	0	0.01	0.03	-0	0.02	-0.01	-0	-0.03	0.07	0.22	-0.06	-1	1

TARGET variable doesn't have any noticeable correlation with Family, Income, Employment, Rating information

Regional Ratings and External scores don't seem to be associated with any other numerical feature

There is a weak but positive correlation between AGE_YEARS & EMPL_YEARS

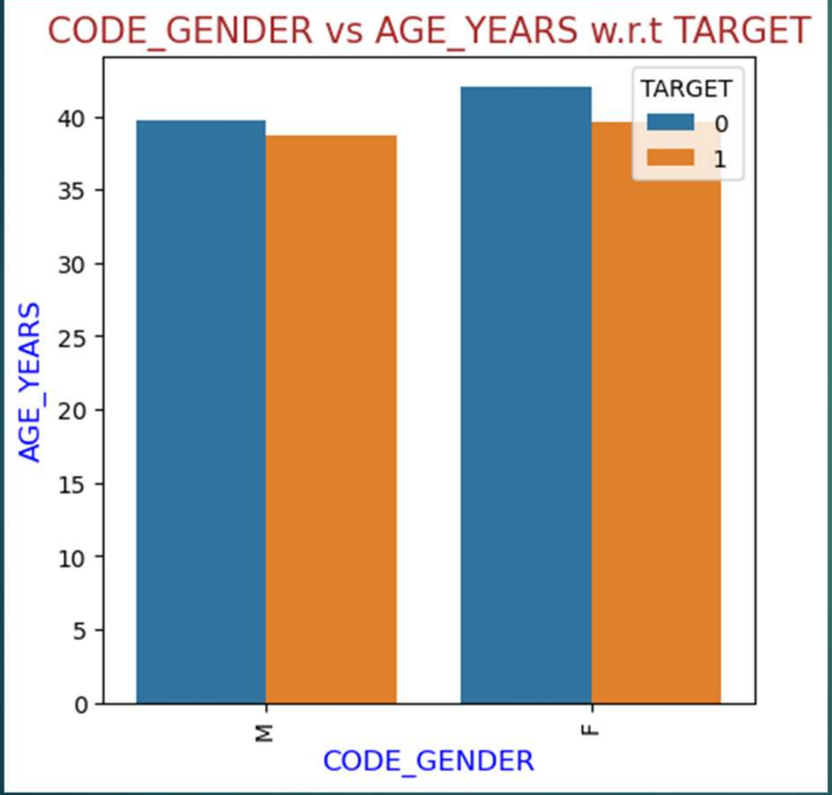
CNT_CHILDREN and CNT_FAM_MEMBERS have a strong positive correlation

Respective ANNUITY and CREDIT variables are also associated with each other positively

Previous loan term CNT_PAYMENT has a stronger correlation with respective CREDIT but weaker with ANNUITY

Income i.e. AMT_TOTAL_INCOME has a fairly positive association with present CREDIT and ANNNUITY variables

Data Analysis & Plotting Graphs
>>> Bivariate Analyses - Categorical:

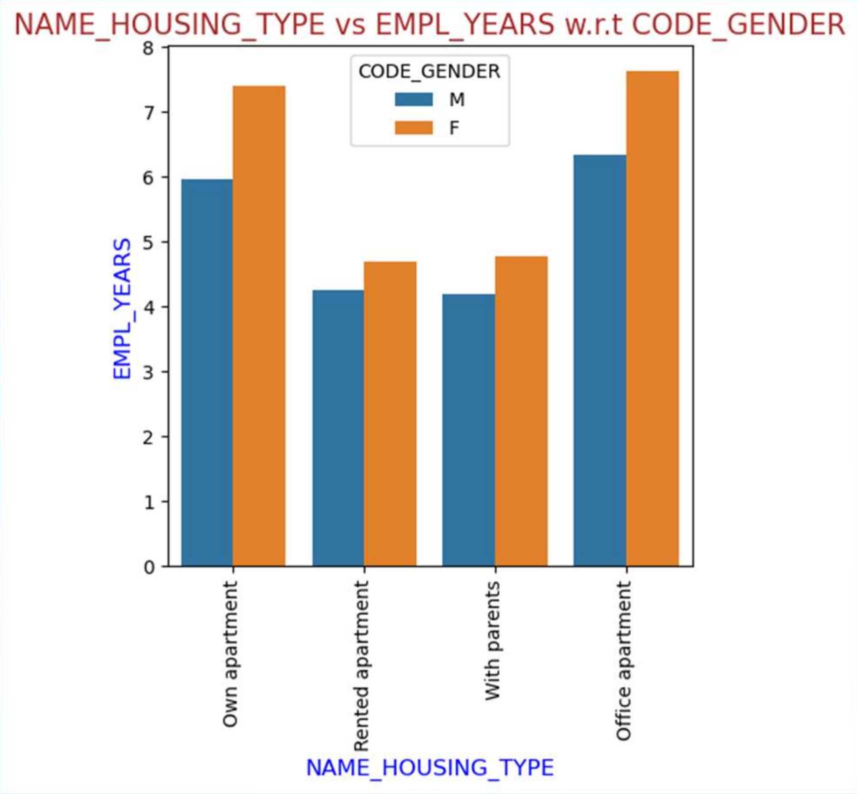


Female clients who are more than 40 years have a little lower probability of repaying their loans as compared to under 40.

Probability of male applicants to repay their loans, regardless of their age, is almost same.

Disclaimer:
TARGET = 1 when a client has difficulty repaying loan
TARGET = 0 for other cases

Data Analysis & Plotting Graphs
>>> Bivariate Analyses - Categorical:

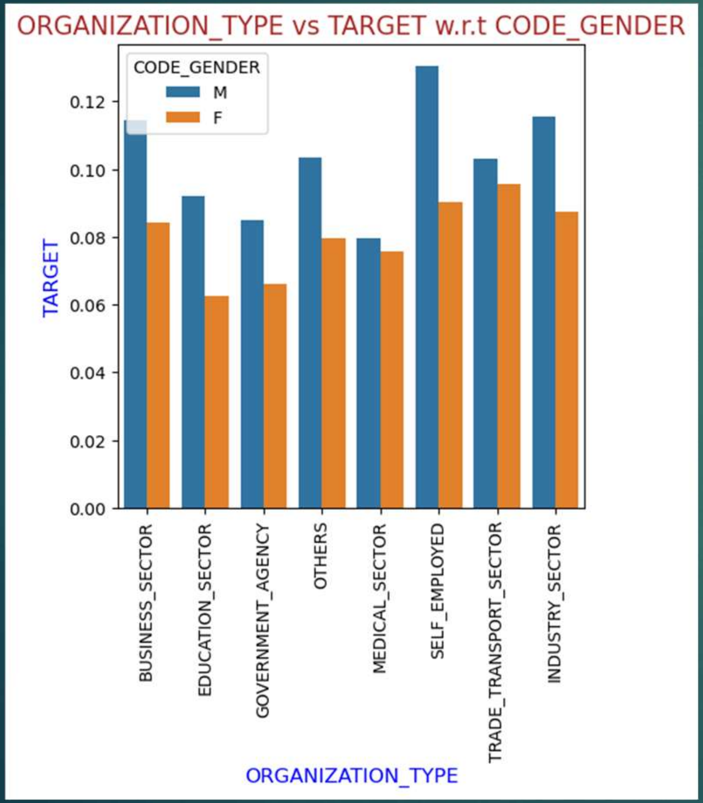


Most number of applicants are female clients with 6 years or more working experience with their own apartment.

Least number of clients are males with around 4 to 5 years of work experience living with parents or in a rented apartment.

Data Analysis & Plotting Graphs

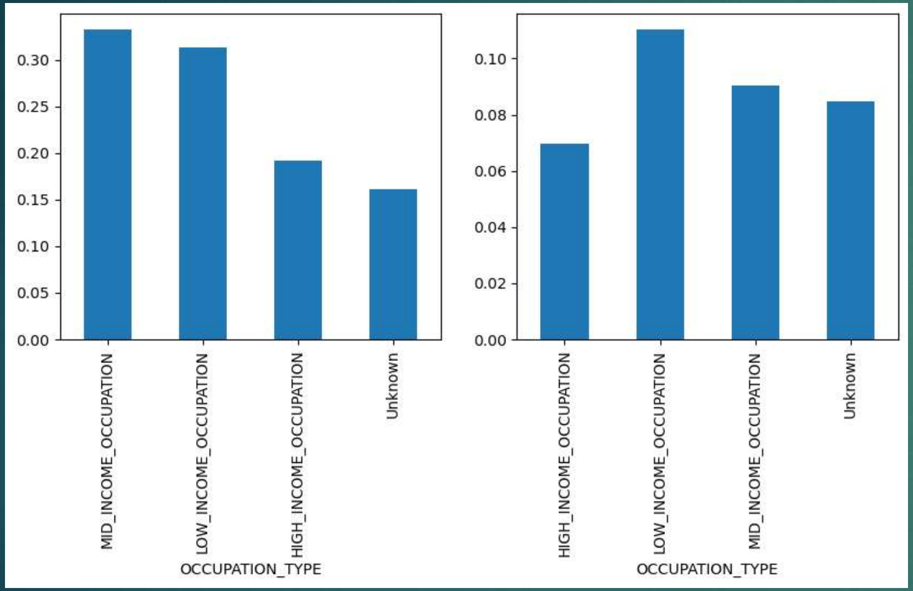
>>> Bivariate Analyses - Categorical:



Applicants who are either running a business or are self employed, have higher number of males who are capable of repaying their loans.

Data Analysis & Plotting Graphs

>>> Bivariate Analyses - Numerical-Categorical:



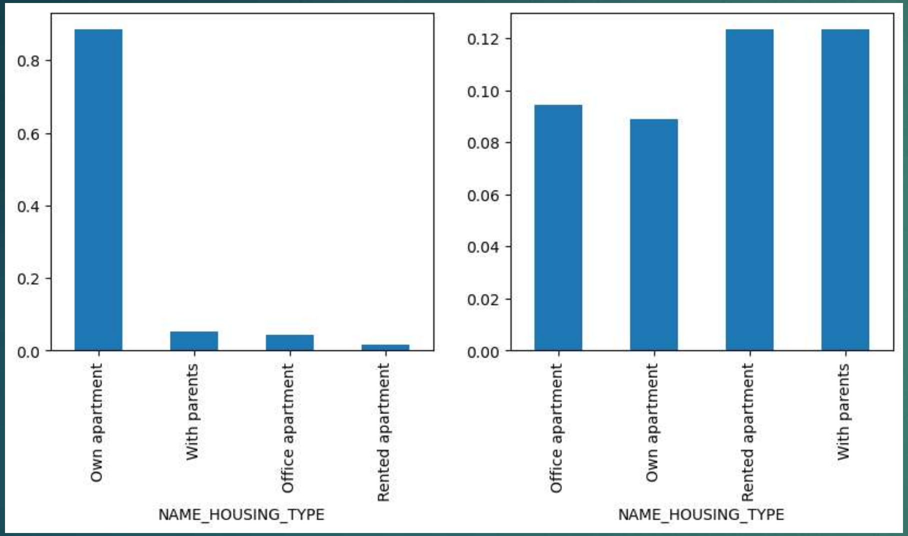
Clients in high income occupation are relatively low in numbers, but have more chances to repay a loan

Applicants from low income group are more likely to default.

Disclaimer:
TARGET = 1 when a client has difficulty repaying loan
TARGET = 0 for other cases

Data Analysis & Plotting Graphs

>>> Bivariate Analyses - Numerical-Categorical:



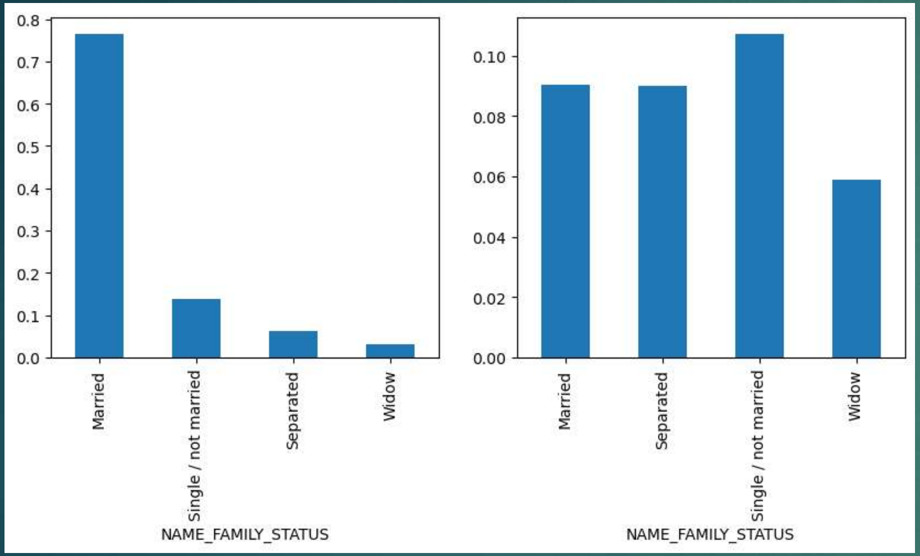
Clients with own house whether it is their own apartment, or office apartment, have higher chances of repaying a loan

Number of clients with their own apartment is the highest, so there a quite good chance for the lending company to recover loan.

Disclaimer:
TARGET = 1 when a client has difficulty repaying loan
TARGET = 0 for other cases

Data Analysis & Plotting Graphs

>>> Bivariate Analyses - Numerical-Categorical:



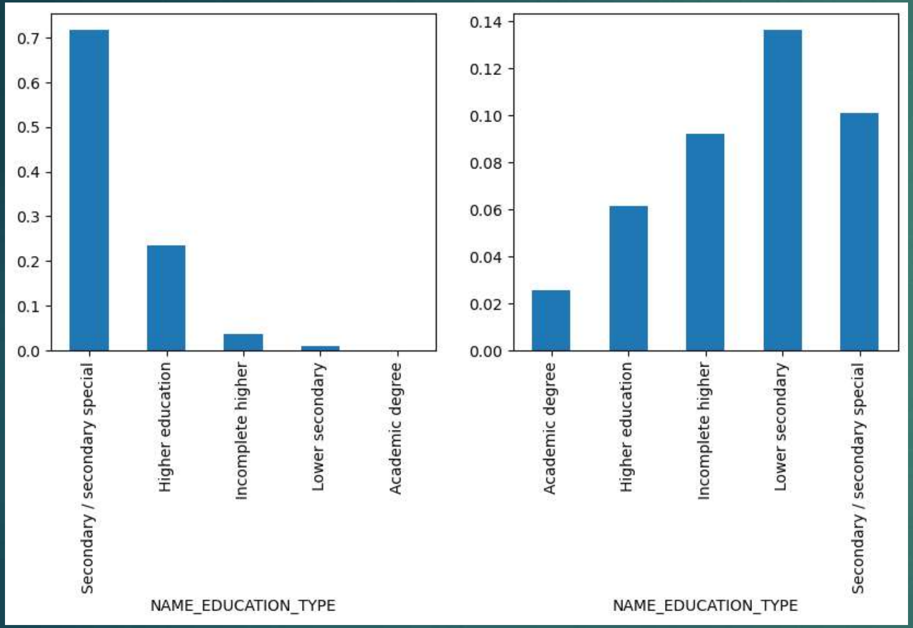
Widow clients have higher probability of repaying a loan, but they are also least in numbers

Married clients are highest in numbers, but not potentially the best in repaying loan

Disclaimer:
TARGET = 1 when a client has difficulty repaying loan
TARGET = 0 for other cases

Data Analysis & Plotting Graphs

>>> Bivariate Analyses - Numerical-Categorical:



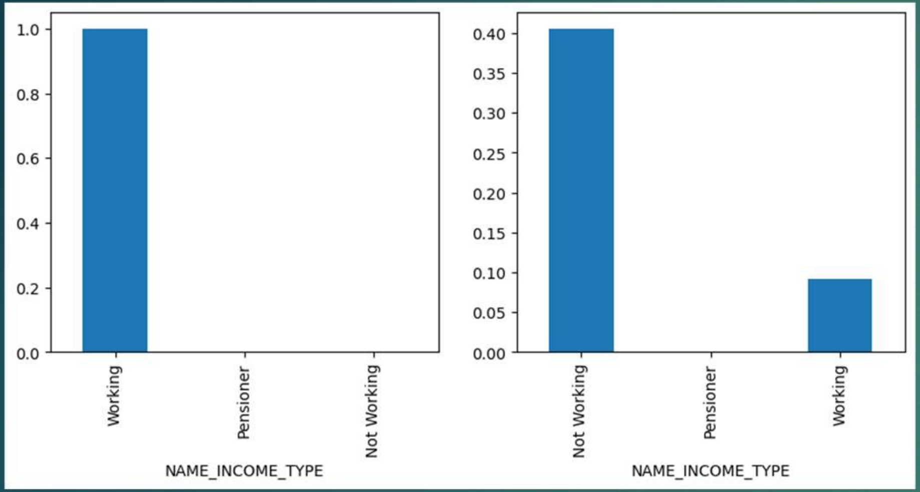
Applicants with an academic degree are the least in numbers, but seem to be most worthy of a loan

Clients with secondary education are highest in applicants, however they are not a very good fit for a loan approval

Disclaimer:
TARGET = 1 when a client has difficulty repaying loan
TARGET = 0 for other cases

Data Analysis & Plotting Graphs

>>> Bivariate Analyses - Numerical-Categorical:



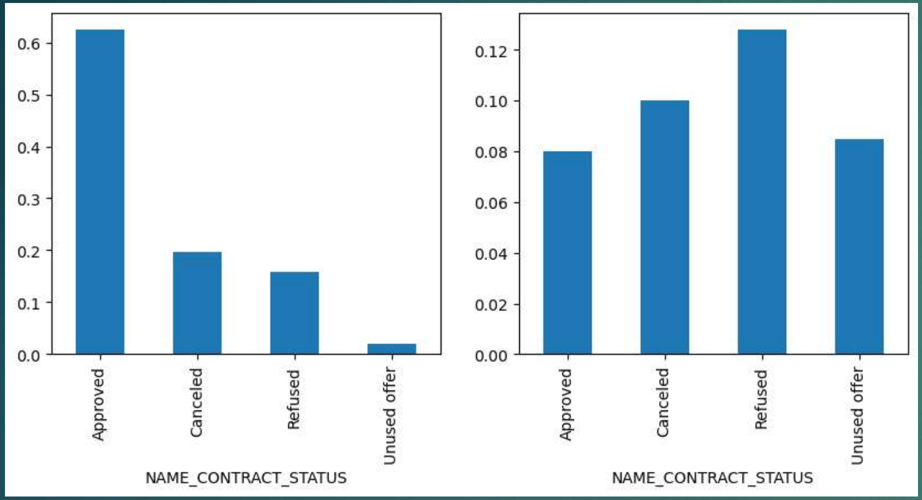
Working applicants are the highest in numbers and also seem to be relatively better worthy of loan

Clients not employed are the one among the least in numbers, but the worst to get a loan approval

Disclaimer:
TARGET = 1 when a client has difficulty repaying loan
TARGET = 0 for other cases

Data Analysis & Plotting Graphs

>>> Bivariate Analyses - Numerical-Categorical:

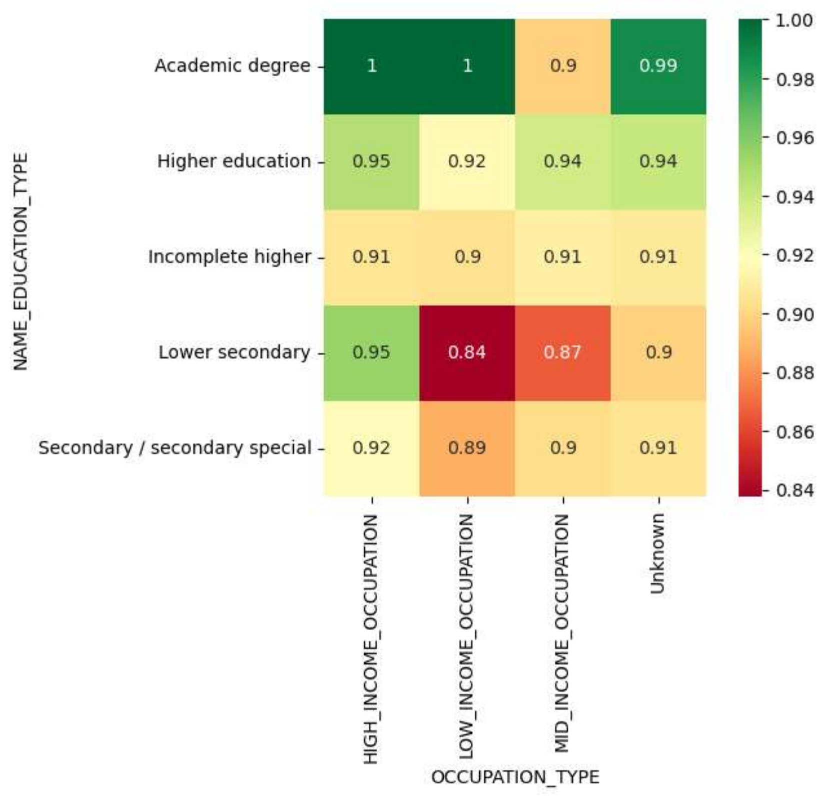


The clients whose loan was approved in last application, are more likely to get another loan approved.

Even the clients who had not used the approved loan offer last time, are loan worthy this time.

Disclaimer:
TARGET = 1 when a client has difficulty repaying loan
TARGET = 0 for other cases

Data Analysis & Plotting Graphs >>> Multivariate Analysis:

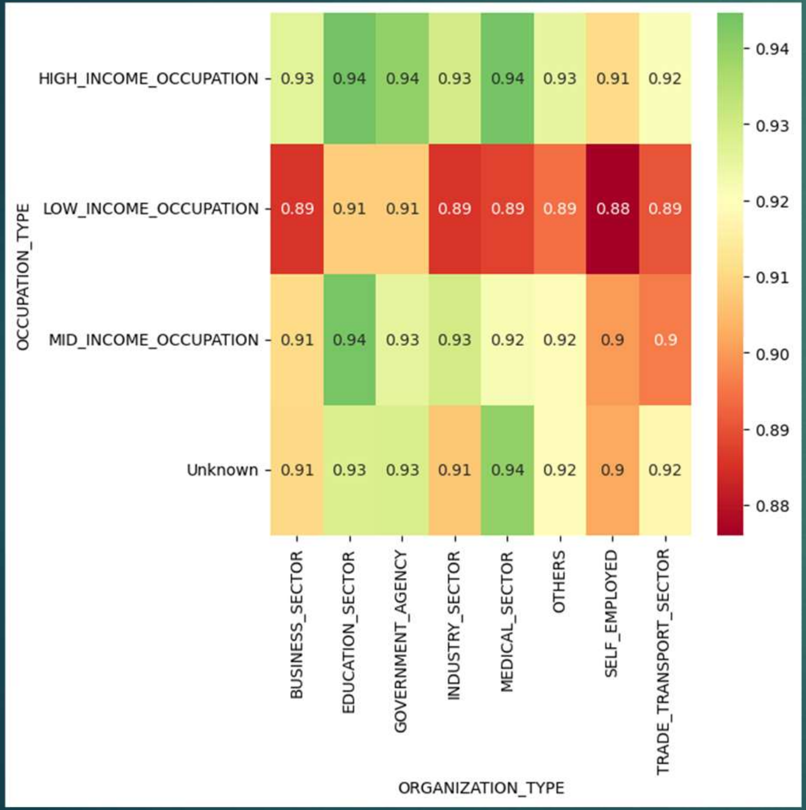


Clients with an academic degree have higher positive correlation regardless of occupation type

Worst case is the clients with lower secondary education and low or mid income occupation

Data Analysis & Plotting Graphs

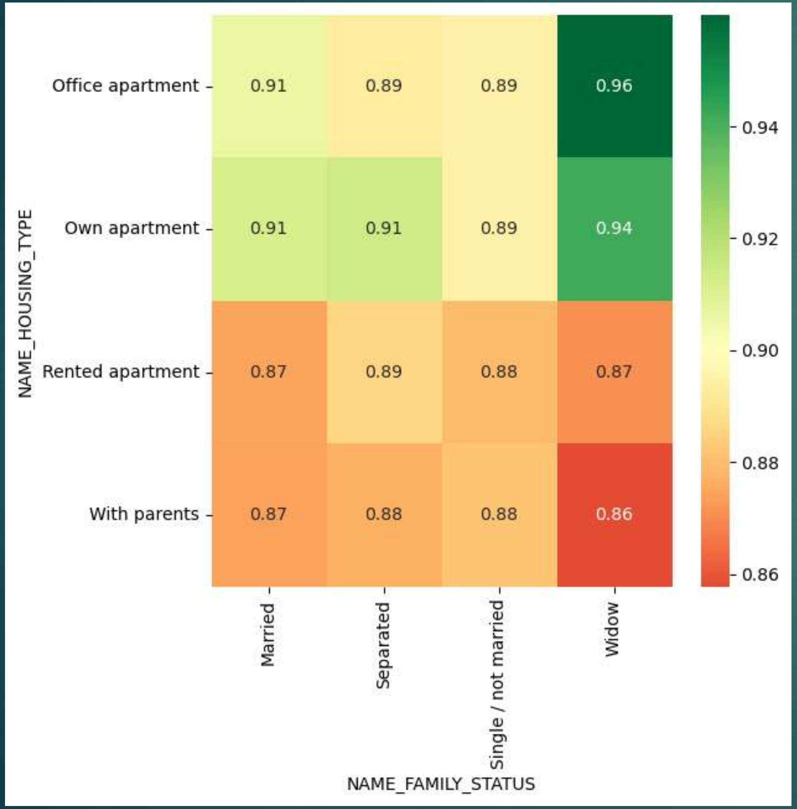
>>> Multivariate Analysis:



Regardless of organization type, clients in low income groups are more likely to default

Clients working in Education, Government, Industrial and Medical sectors are the ones who are most likely to repay their loan

Data Analysis & Plotting Graphs >>> Multivariate Analysis:



In general, the clients either staying in a rented apartment or with parents are the least worthy of a loan

Exceptions among these is the clients staying in own or office apartment who are separated or single

The worst case of a client not worthy of a loan, would be the widow clients staying with parents

Conclusions Drawn

>>> Variables indicating potential default cases:

- Education Type: Clients with secondary education are not very likely in their loan repayment.

- Occupation Type: Applicants from low income group such as security/cooking/sales staff and drivers are more likely to default.

- Gender: There is a lower probability of male clients repaying a loan as compared to female clients. Female clients who are more than 40 years have a little lower probability of repaying their loans as compared to under 40.

Conclusions Drawn

>>> Variables indicating potential default cases:

- ❑ Family Status: There is a lower probability of repaying a loan if a client is single, separated, or a widow as compared with the applicants who are married.
- ❑ Housing Type: There is a lower probability of repaying a loan where a client is staying with parents, or in a rented apartment as compared to a client having own or office apartment.
- ❑ Realty Ownership: A client not owning a property has lower chances of repaying a loan.

Conclusions Drawn

>>> Variables indicating potential default cases:

- ❑ Contract Type: A client who has applied for a revolving loan is less likely to repay loan as compared to an applicant who applied for a cash loan.

- ❑ Income Type: Quite naturally, the clients who are not working such as pensioners are going to have less probability of repaying their loans as compared to working applicants.

- ❑ Organization Type: The clients who are working in trade, transport or industry sectors have quite low probability of repaying their loans as compared to business class or self employed clients.

Conclusions Drawn

>>> Correlations for TARGET variable:

- Clients with an Academic degree or Higher Education have higher positive correlation with TARGET, regardless of Occupation type.
- Worst correlation between TARGET and Education level is with the clients who have obtained lower secondary education and are working in low or mid income Occupation types.
- Clients working in Education, Government, Industrial and Medical sectors have a fairly positive correlation with TARGET, regardless of Occupation type.
- Worst correlation between TARGET and Occupation type is with the clients who are working in LOW income occupations such as Laborers, Waiters, Cooks, Security staffs, regardless of Organization Type.

Conclusions Drawn

>>> Correlations for TARGET variable:

- Clients staying in own or office apartment have higher positive correlation with TARGET, regardless of Family status.
- The worst case between TARGET and Housing type is such a widow client staying with parents.

Thank You