Exploratory Data Analysis for Banking and financial services

Case Study

28 Feb 2023

By -
Yasaswi Racha

Table of Contents

Business Problem & Objectives

Problem -

To develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Objective -

- The company wants to understand the driving behind default loan so that company can utilize this knowledge for its portfolio and risk assessment.

- Aim is to identify patterns which indicate if the clients who are taking loan are capable of paying their installments. If not taking the required actions and rejecting the loan amount for the candidates who are not capable of repaying the loan. Identification of such applicants using EDA is the aim of this case study.

Data Importing

- Importing different libraries like numpy, matplotlib, Pandas, etc
- Importing data Application data, previous data and column data

1	application_data	SK_ID_CURR	ID of loan in our sample	Unnamed: 4	
0	2	application_data	TARGET	Target variable (1 - client with payment diffi...	NaN
1	5	application_data	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving	NaN
2	6	application_data	CODE_GENDER	Gender of the client	NaN
3	7	application_data	FLAG_OWN_CAR	Flag if the client owns a car	NaN
4	8	application_data	FLAG_OWN_REALTY	Flag if client owns a house or flat	NaN

Fig Column Data

Summary of the Data

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_APPLICATION	AMT_CREDIT	WEEKDAY_APPR_PROCESS_START	HOUR_APPR_PROCESS_START
4	1784265	202054	Cash loans	337500.0	404055.0	THURSDAY	9
5	1383531	199383	Cash loans	315000.0	340573.5	SATURDAY	8
19	1173070	199178	Cash loans	45000.0	49455.0	SATURDAY	16
51	2664403	163660	Cash loans	67500.0	82611.0	SATURDAY	10
85	1981960	306707	Cash loans	225000.0	269550.0	WEDNESDAY	11

5 rows × 22 columns

Previous Application Data

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY
0	100002	1	Cash loans	M	N	Y
1	100003	0	Cash loans	F	N	N
2	100004	0	Revolving loans	M	Y	Y
3	100006	0	Cash loans	F	N	Y
4	100007	0	Cash loans	M	N	Y

5 rows × 29 columns

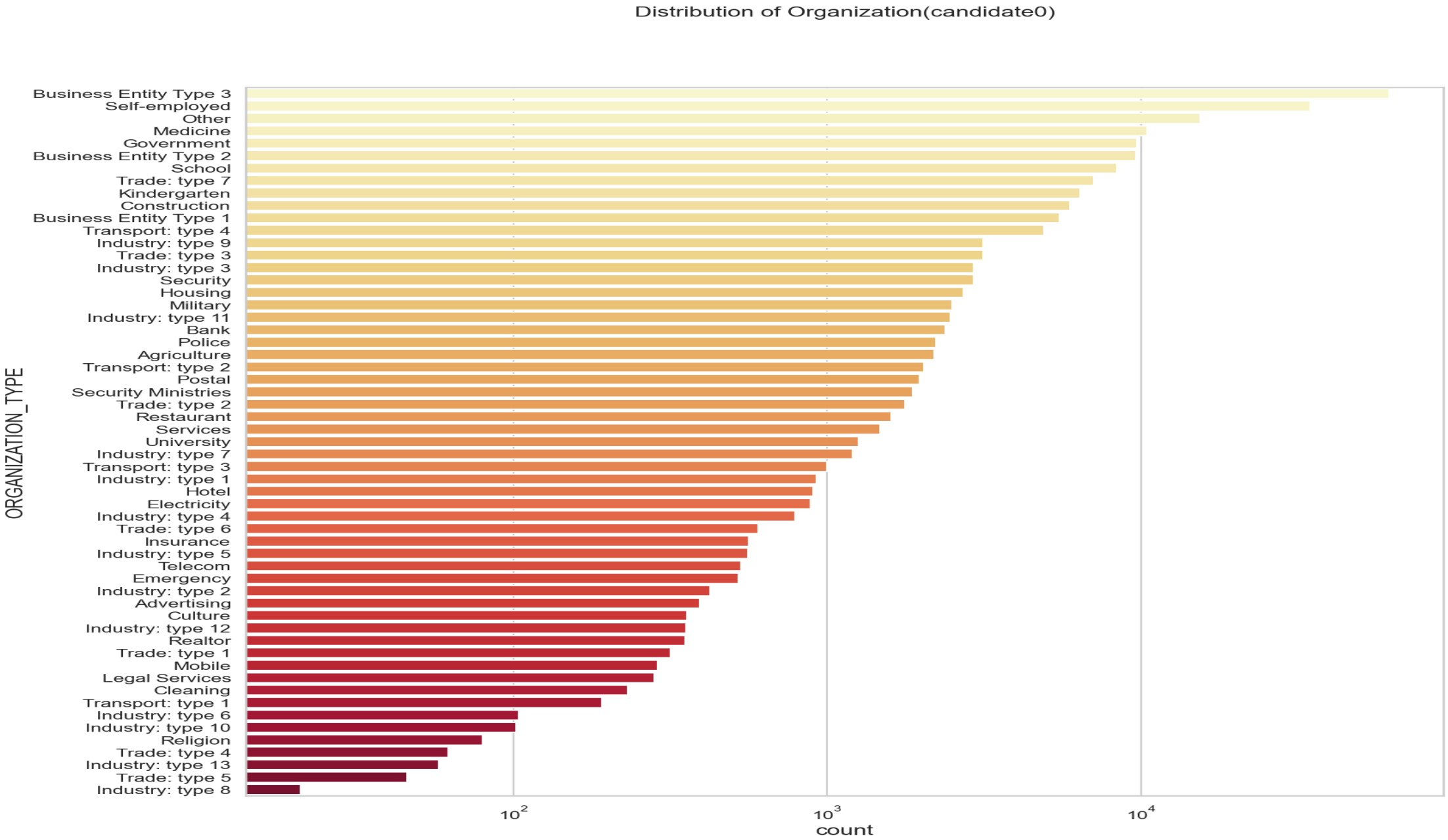
Application Data

Data Cleaning and Data Manipulation of Application Data

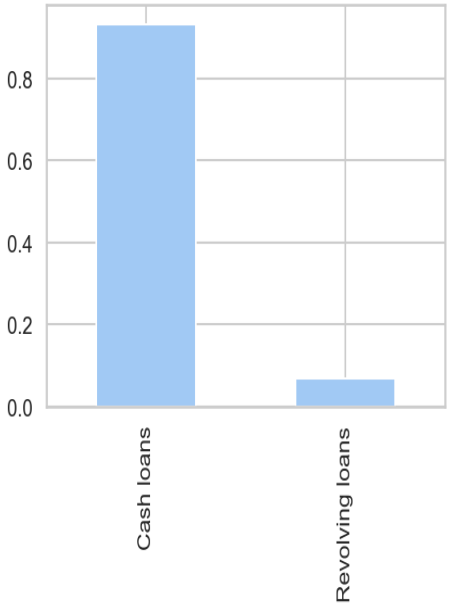
- Dropping the first column of application data as it does not help with any data
- Cleaning the missing data
- Removing 64 columns of null data
- Checking the columns having less null percentage
- Filling missing values with median as it cannot be empty
- Checking and removing rows as well as columns having null values which have more or equal to 30% in the data
- Checking Gender column for the number of females and males data and updating the missing information with of CODE_GENDER with F
- Converting all variable into numeric in the dataset
- Differentiating and dividing the dataset into two datasets of client with payment difficulties as candidate1 and others as candidate0

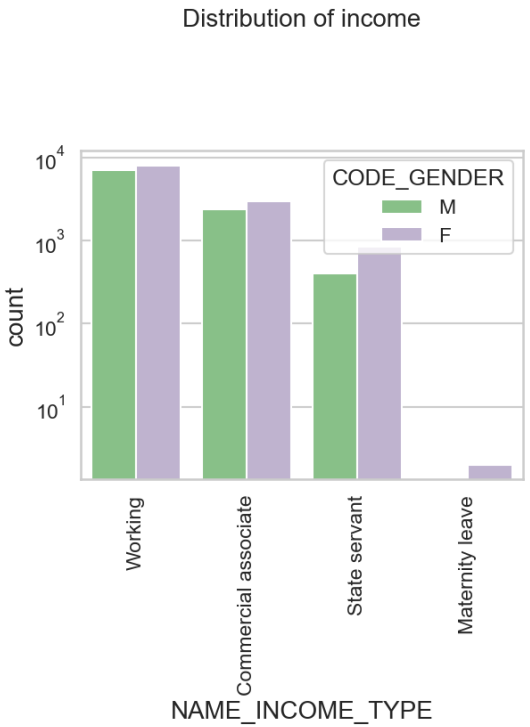
Univariate Analysis (clients who can pay loan)

- According to this Females are having more number of credits than male.
- For income type working, commercial associate, and State Servant the number of credits are higher than others.
- Less number of credits for income type student, pensioner, Businessman and Maternity leave.
- In distribution graph clients which have applied for credits are from most of the organization type Business entity Type 3 , Self employed, Other , Medicine and Government.
- Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4.

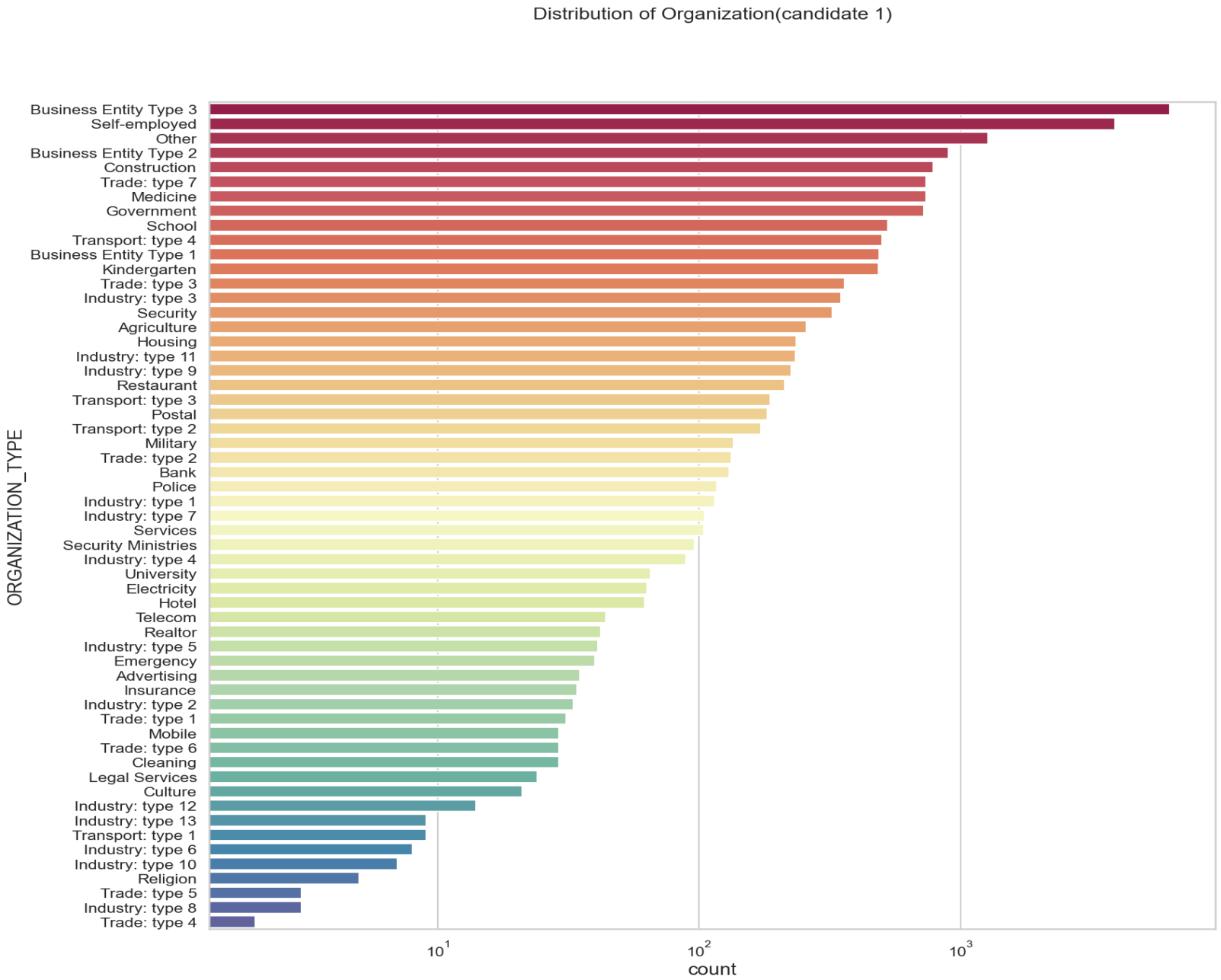

Distribution of Organization(candidate0)
ORGANIZATION_TYPE
Business Entity Type 3
Self-employed
Other
Medicine
Government
Business Entity Type 2
School
Trade: type 7
Kindergarten
Construction
Business Entity Type 1
Transport: type 4
Industry: type 9
Trade: type 3
Industry: type 3
Security
Housing
Military
Industry: type 11
Bank
Police
Agriculture
Transport: type 2
Postal
Security Ministries
Trade: type 2
Restaurant
Services
University
Industry: type 7
Transport: type 3
Industry: type 1
Hotel
Electricity
Industry: type 4
Trade: type 6
Insurance
Industry: type 5
Telecom
Emergency
Industry: type 2
Advertising
Culture
Industry: type 12
Realtor
Trade: type 1
Mobile
Legal Services
Cleaning
Transport: type 1
Industry: type 6
Industry: type 10
Religion
Trade: type 4
Industry: type 13
Trade: type 5
Industry: type 8
10^2
10^3
10^4
count

Univariate Analysis (clients with difficulty to repay loan)

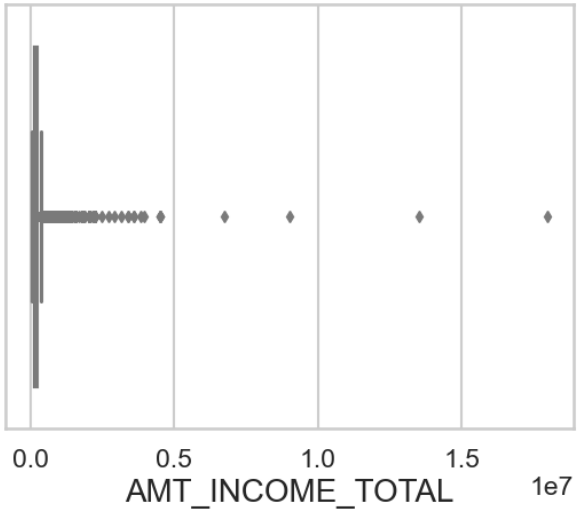




- For income type working, commercial associate, and State Servant the number of credits are higher than other i.e. Maternity leave
- Females are having more number of credits than male.
- For contract type cash loans is having higher number of credits than Revolving loans contract type
- In distribution of organization clients which have applied for credits are from most of the organization type Business entity Type 3 , Self employed , Other , Medicine and Government
- Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4
- Same as type 0 in distribution of organization type

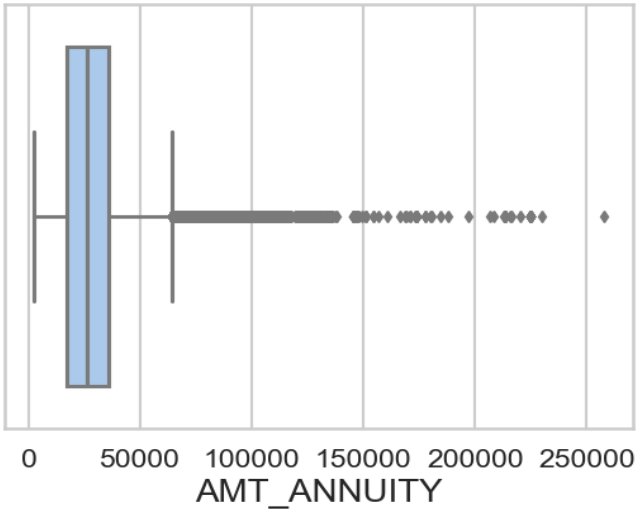

Distribution of Organization(candidate 1)
ORGANIZATION_TYPE
Business Entity Type 3
Self-employed
Other
Business Entity Type 2
Construction
Trade: type 7
Medicine
Government
School
Transport: type 4
Business Entity Type 1
Kindergarten
Trade: type 3
Industry: type 3
Security
Agriculture
Housing
Industry: type 11
Industry: type 9
Restaurant
Transport: type 3
Postal
Transport: type 2
Military
Trade: type 2
Bank
Police
Industry: type 1
Industry: type 7
Services
Security Ministries
Industry: type 4
University
Electricity
Hotel
Telecom
Realtor
Industry: type 5
Emergency
Advertising
Insurance
Industry: type 2
Trade: type 1
Mobile
Trade: type 6
Cleaning
Legal Services
Culture
Industry: type 12
Industry: type 13
Transport: type 1
Industry: type 6
Industry: type 10
Religion
Trade: type 5
Industry: type 8
Trade: type 4
10^1
10^2
10^3
count

Univariate variables analysis for candidate-0



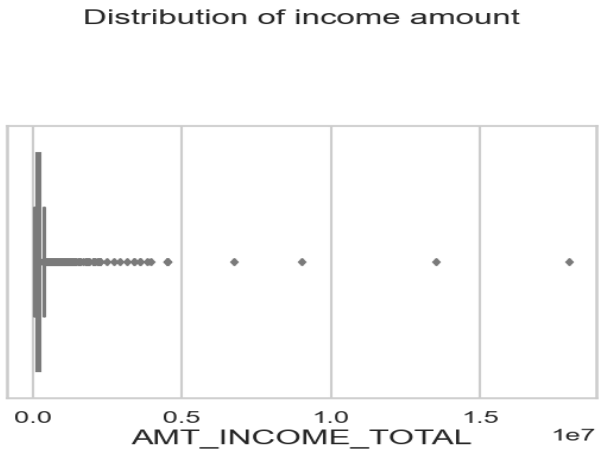
- Some outliers are noticed in income amount
- The third quartiles is very slim for income amount

- Some outliers are noticed in credit amount.
- The first quartile is bigger than third quartile for credit

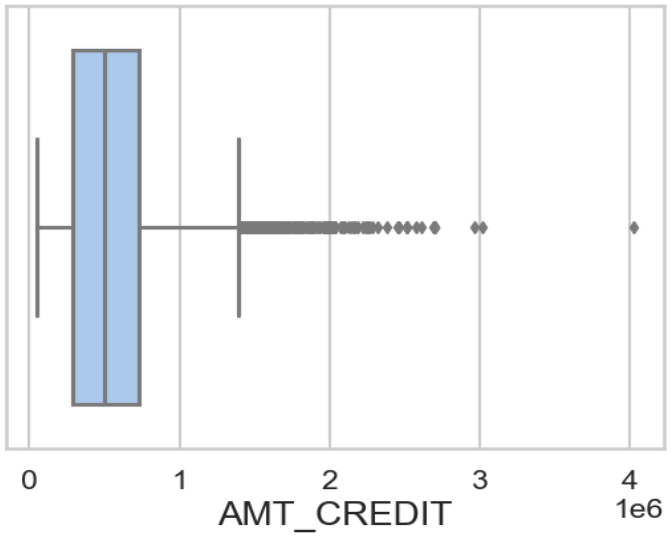


- Some outliers are noticed in annuity amount.
- The first quartile is bigger than third quartile for annuity amount showing first quartile have more clients

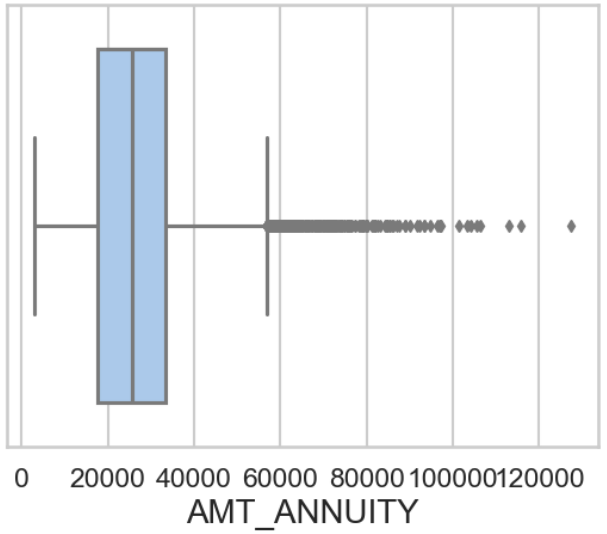
Univariate variables analysis for candidate-1



- Some outliers are noticed in income amount
- Most of the clients of income are present in first quartile

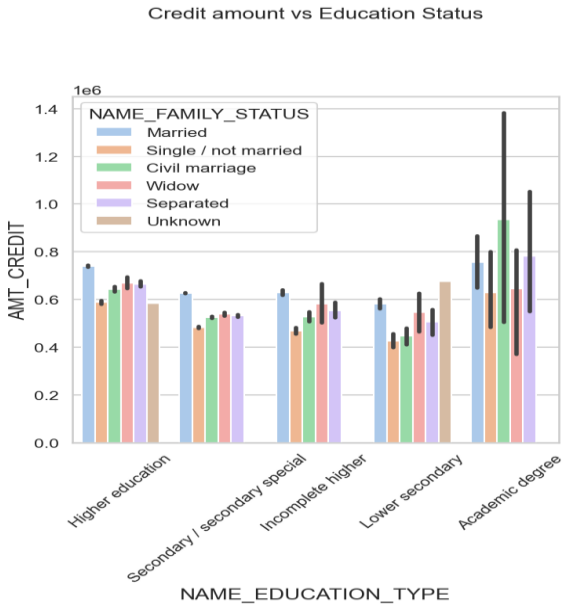


- Some outliers are noticed in credit amount
- The first quartile is bigger than third quartile for credit amount so most of the credits of clients are present in the first quartile

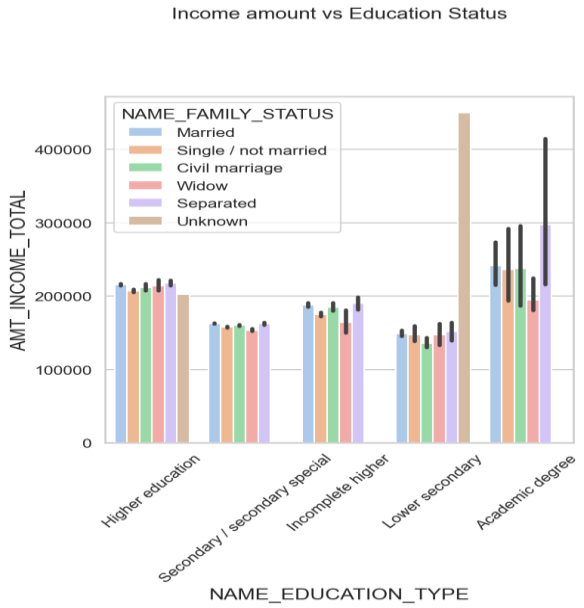


- Some outliers are noticed in annuity amount
- The first quartile is bigger than third quartile for annuity amount which means most of the annuity clients are from first quartile

Bivariate analysis for candidate-0

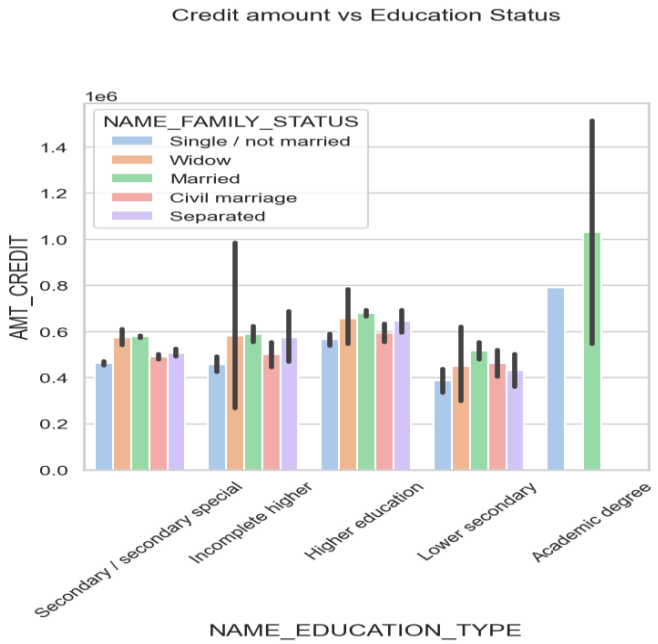


Family status of civil marriage, marriage and separated of Academic degree education are having higher number of credits than others

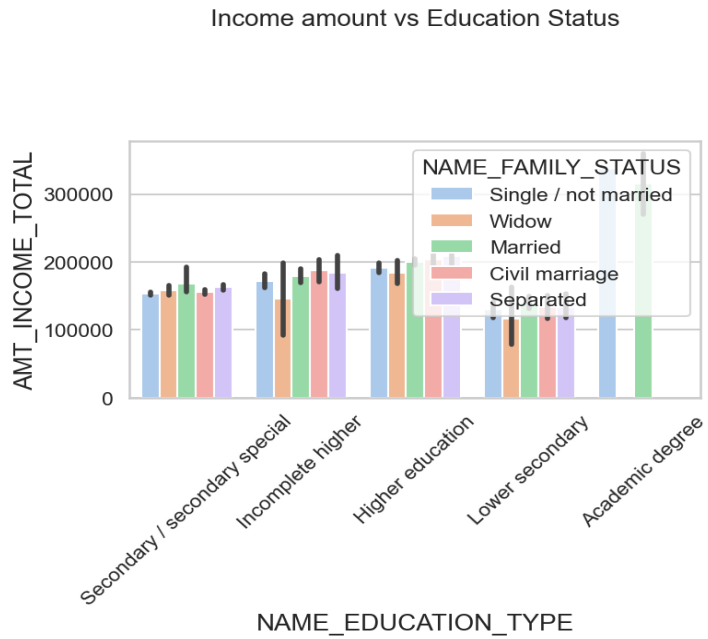


Higher education the income amount is mostly equal with family status. Less outlier are having for Academic degree but there income amount is little higher that Higher education. Lower secondary of civil marriage family status are have less income amount than others

Bivariate analysis for candidate-1

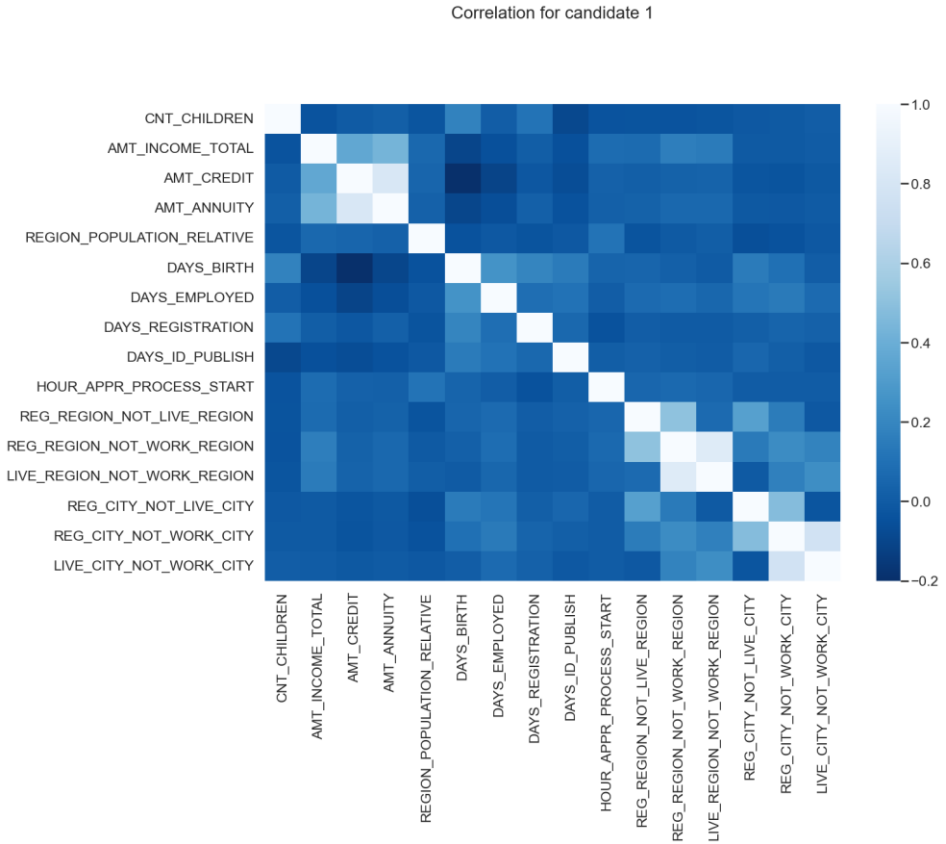


Family status of civil marriage, marriage and separated of Academic degree education are having higher number of credits than others. The outliers are from Higher education and Secondary. Most of people who are married have academic degree.



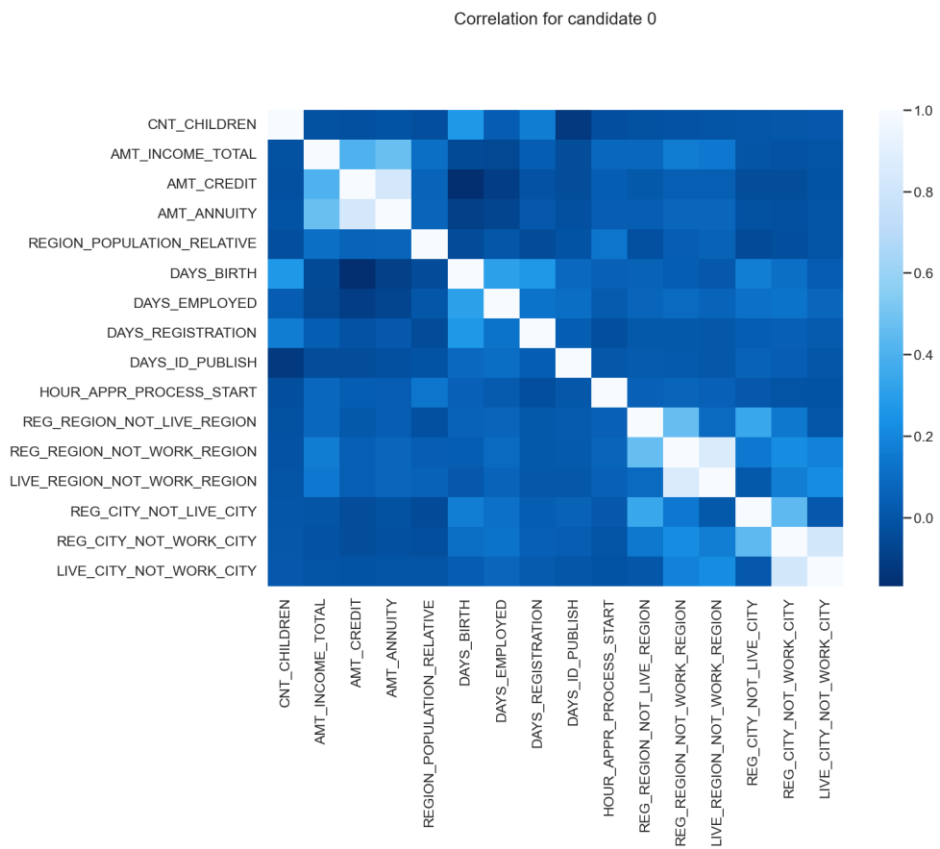
Higher education the income amount is mostly equal with family status. Less outlier are having for Academic degree but there income amount is little higher that Higher education. Lower secondary are have less income amount than others.

Correlation for both candidates who cannot repay loans



- Credit amount is inversely proportional to the date of birth and number of children client which means Credit amount is higher
- The client's permanent address does not match contact address are having less children and vice-versa
- The client's permanent address does not match work address are having less children and vice-versa
- Credit amount is higher to densely populated area
- The income is also higher in highly populated area

Correlation for both candidates who can repay loans

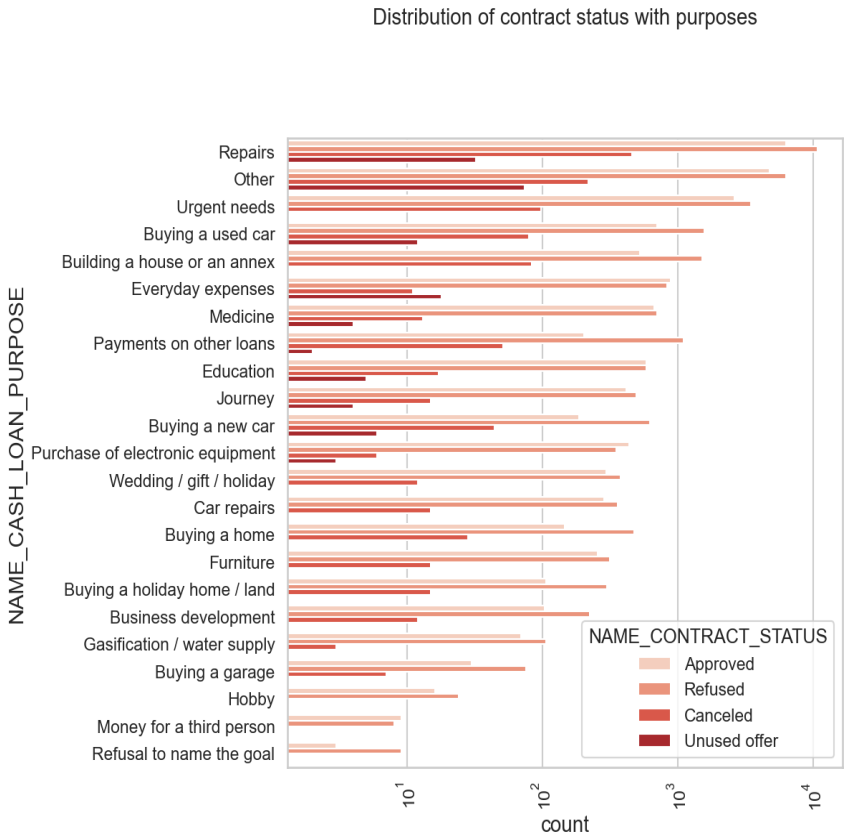


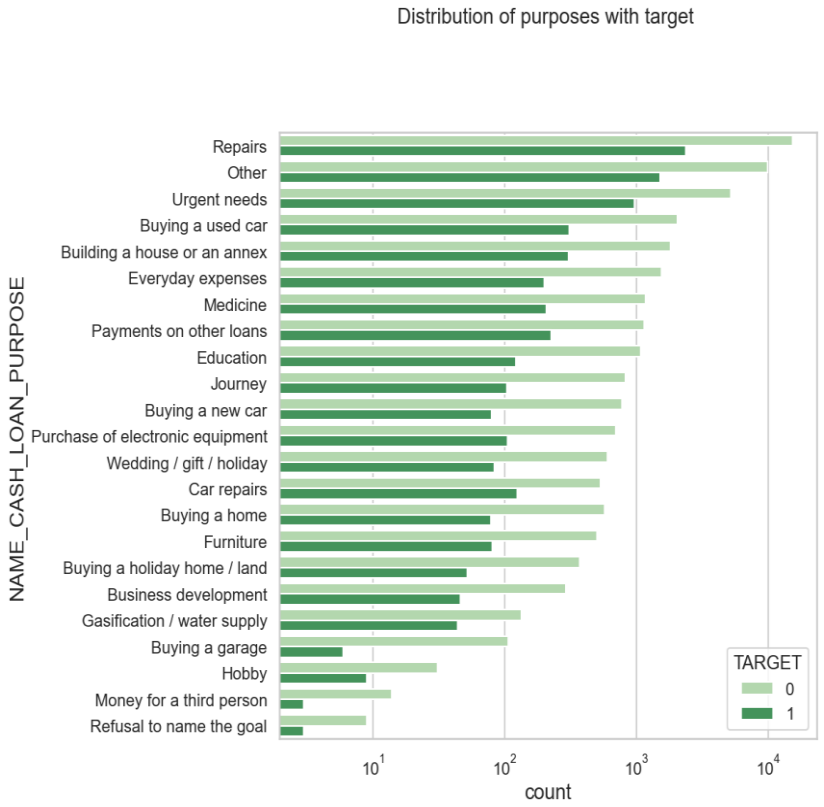
- Credit amount is inversely proportional to the date of birth, which means Credit amount is higher for low age and vice-versa
- Credit amount is inversely proportional to the number of children client have, means Credit amount is higher for less children count client have and vice-versa
- Income amount is inversely proportional to the number of children client have, means more income for less children client have and vice-versa
- Less children client have in densely populated area
- Credit amount is higher to densely populated area
- The income is also higher in highly populated area

Cleaning and Merging Data of previous data

- Cleaning the missing data of previous application dataset
- listing the null values columns and removing 15 columns from the data set
- Joining the Application dataset with previous application dataset
- Renaming the column names after merging
- Removing unwanted columns
- Performing data analysis on the merged dataset

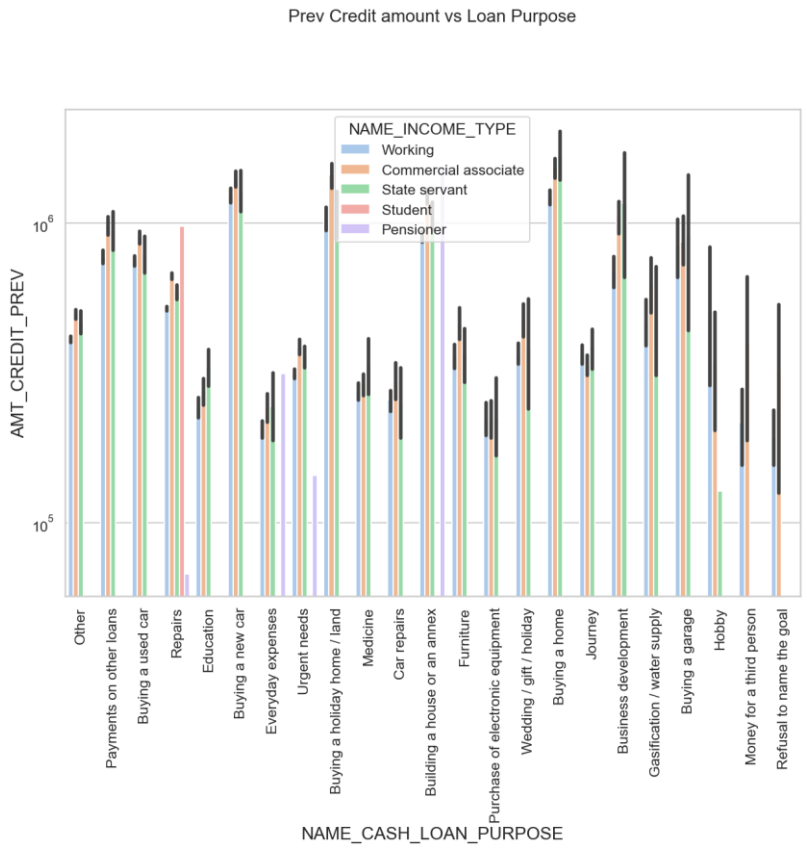
Performing univariate analysis

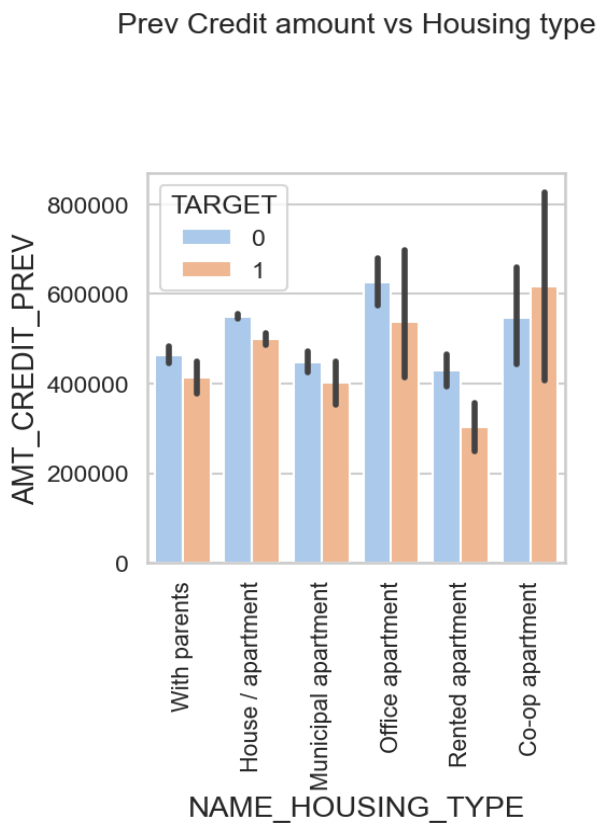




- Most rejection of loans came from purpose repairs
- For education purposes we have equal number of approves and rejection
- Loan purposes with 'Repairs' are facing more difficulties in payment on time. There are few places where loan payment is significant higher than facing difficulties

Performing Bivariate analysis





- The credit amount of Loan purposes like Buying a home and land, buying a new car and Building a house is more.
- Income type of state servants have a significant amount of credit applied
- Housing type, office apartment is having higher credit of target 0 and co-op apartment is having higher credit of target 1.

Conclusion

- Findings say that Banks should focus more on contract type Student ,pensioner and Businessman with housing type other than Co-op apartment for loan repayment as they have more success rate comparatively to others.
- As most of the unsuccessful payments are from income type working , banks should focus more on other income types.
- Purpose of loan for Repair have highest number of unsuccessful payments. So decreasing the loans for Repairs would be better in future.
- Most of the loans repay is from Housing type with parents, hence focusing on that segment will give more success rate for repayment of loans.