Credit EDA Assignment:

Risk Analytics in Banking and Financial Services

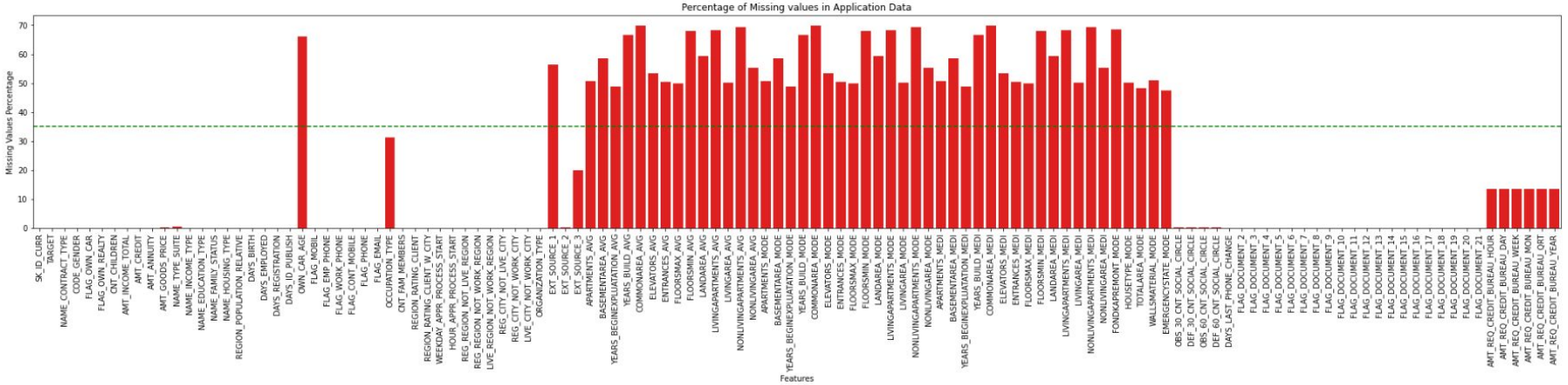
Name of student: Thanh Tra Tran (Rachel)
Email: susutran7575@gmail.com

Business scenario overview:

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because some people can become defaulters => Using EDA to analyse the data to ensure that the applicants capable of repaying the loan are not rejected.
- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. There are two types of risk:
    - If the applicant can repay the loan, then not approving the loan -> lost of business.
    - If the applicant can not repay the loan (default) then approving the loan -> lost cost.

Missing Values

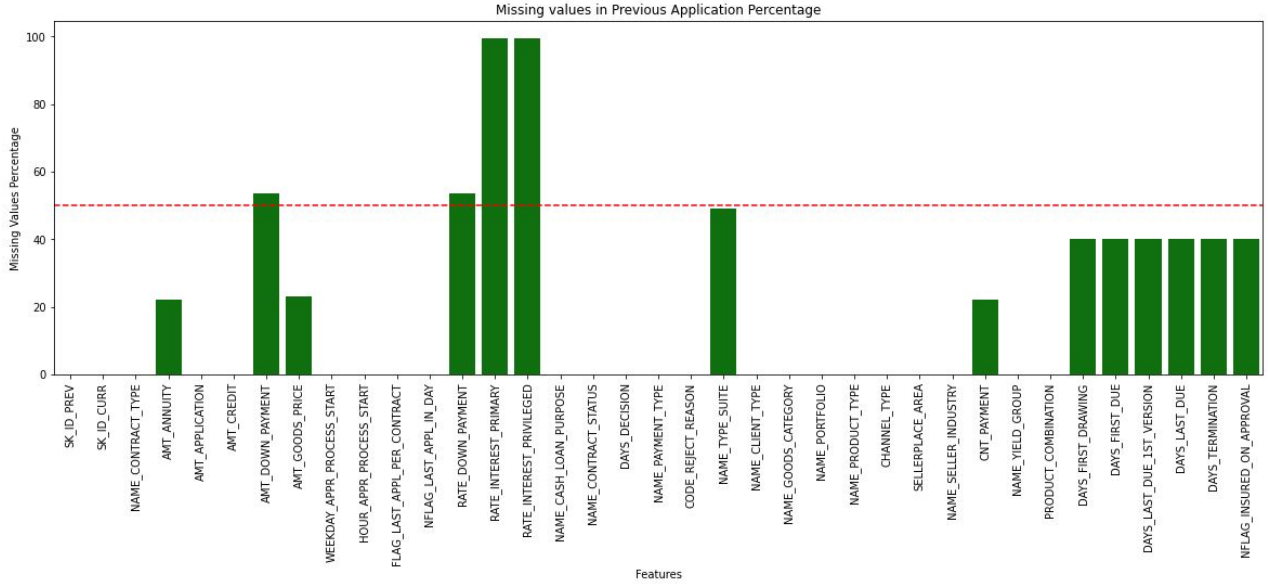
Missing Values on Application Data



Insight:

- 35% missing values will be the cut-off marks. There are 49 columns have missing values more than 35% in Application Data.
- Most of the missing values columns are related to area where the client lives, their assets,...

Missing Values on Previous Application

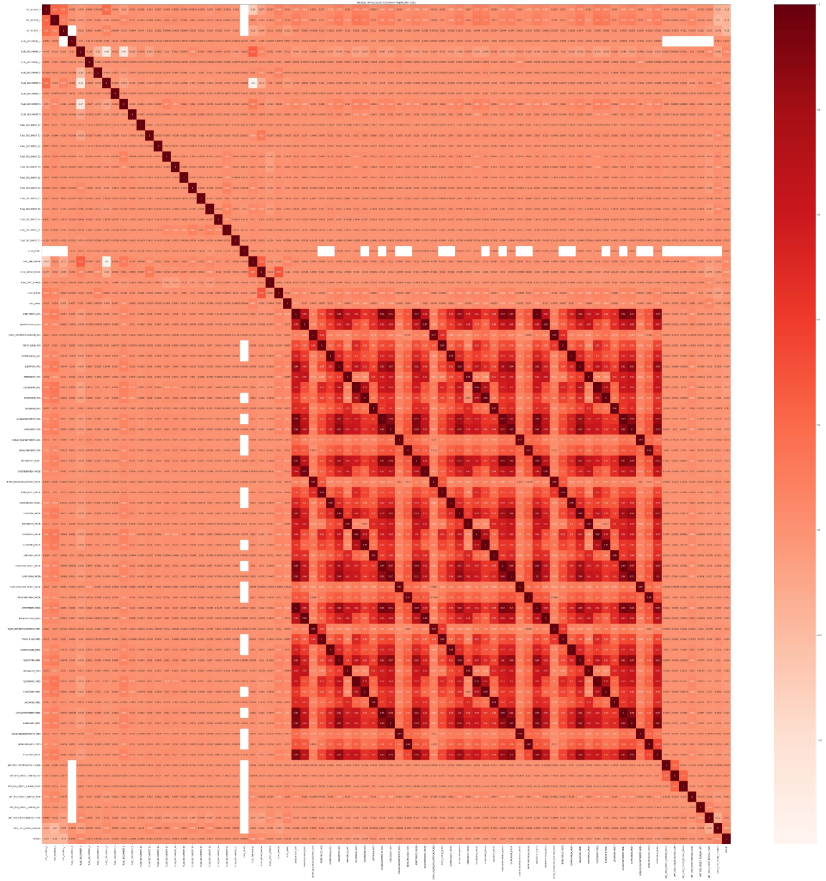


Insight:

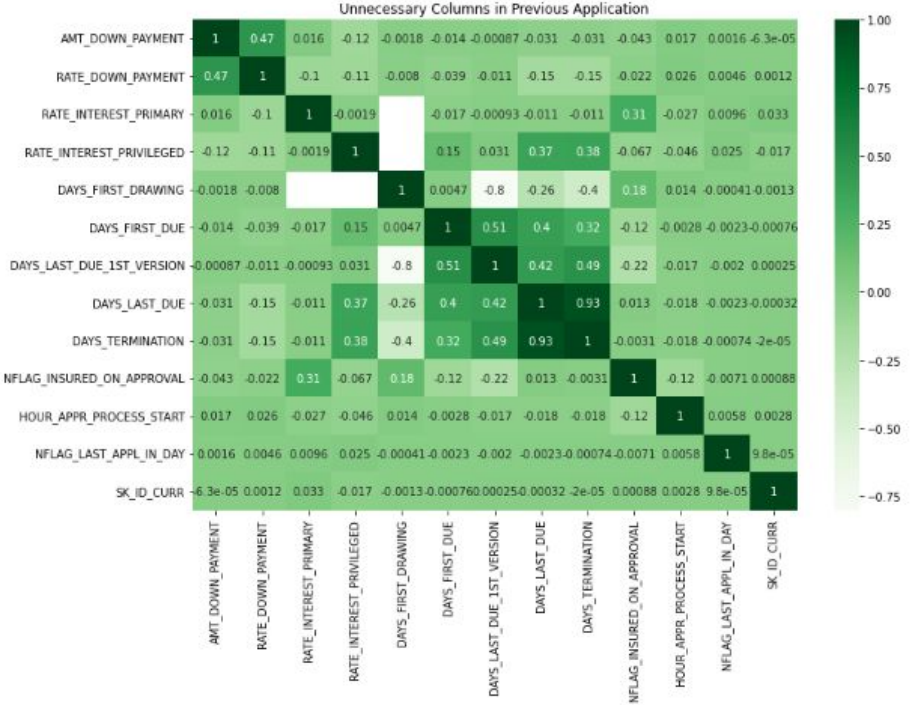
- 45% missing values will be the cut-off marks. There are 5 columns have missing values more than 45% in Previous Application.
- Most of the missing values columns are related to client payment, rate interest,....

Correlation between Target and other missing columns

Unnecessary Columns in Application Data



Unnecessary Columns in Previous Application



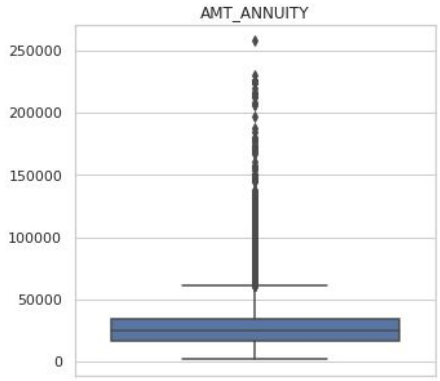
Insight: According to the Heatmap, there is almost no correlation between missing values columns or unnecessary columns and TARGET column so that we can drop or ignore those.

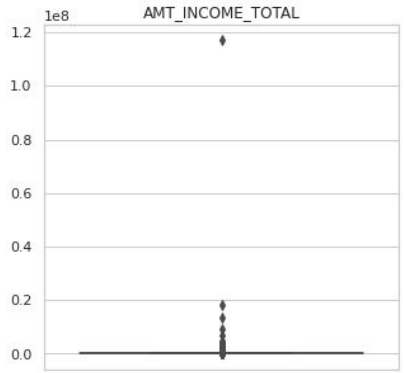
Finding Outlier

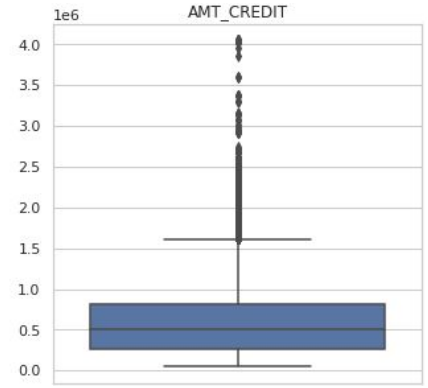
Outlier on Application Data

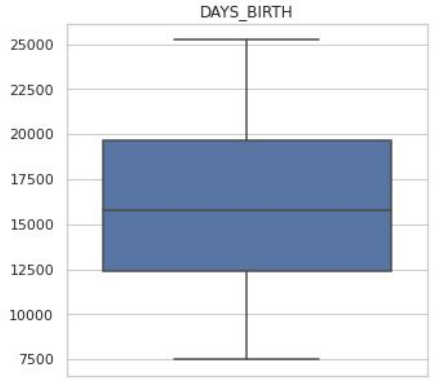
Insight:

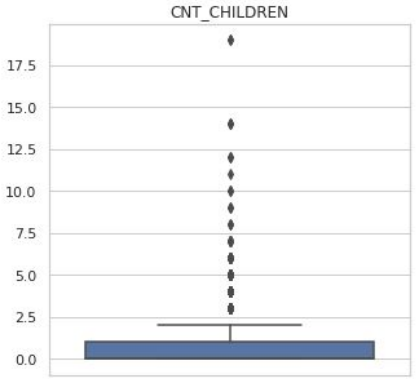
- DAYS_EMPLOYED have outlier value over 350000 days (~958 years).
- AMT_INCOME_TOTAL, CNT_CHILDREN has huge number than the others.
- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE have some number of outliers.
- DAYS_BIRTH has no outliers.

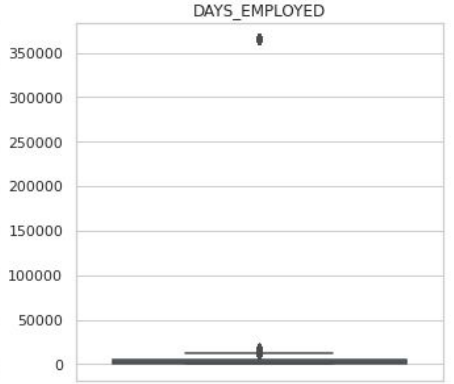




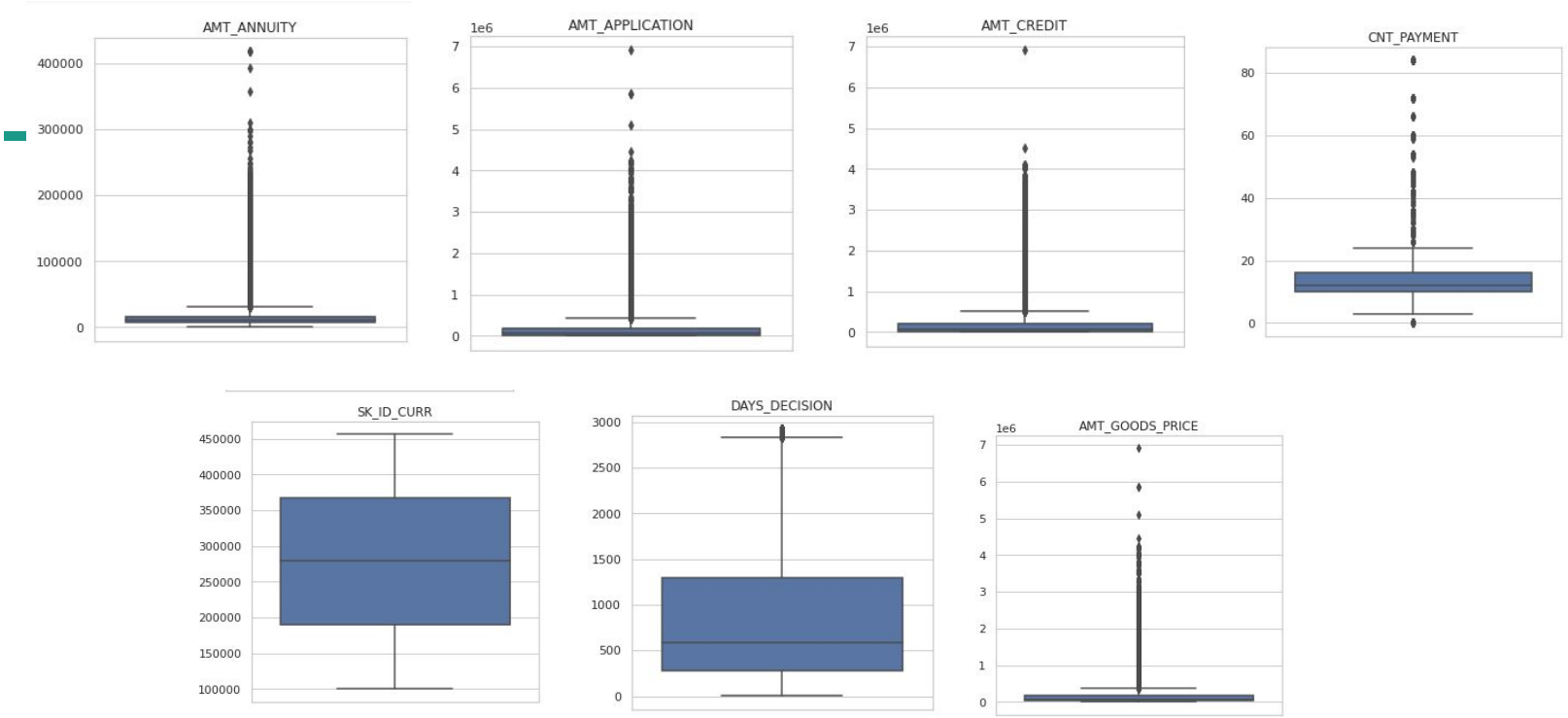








Outlier on Previous Application



Insight:

- AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA, CNT_PAYMENT, DAYS_DECISION have outlier values.
- SK_ID_CURR has no outliers.

Analysis and Insight

Analyze Data on Application Data

Target Variable: **TARGET column**

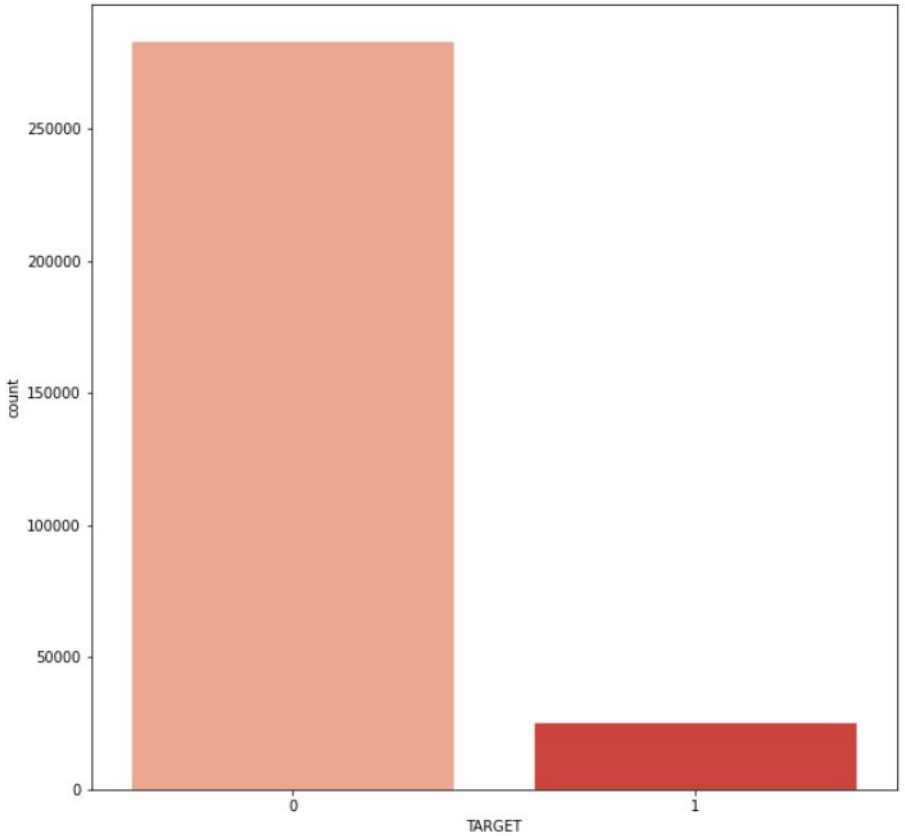
1 - The client with payment difficulties: they had late payment more than X days on at least one of the first Y installments of the loan.

0 - All other cases when the payment is paid on time.

Insight:

- Ratios in percentage with respect to clients with payment difficulties and all other cases in dataset are: 91.93% and 8.07%
- The imbalance is very high.

=> Very few applications have repaid the loan on time, so the risk of clients not being able to repay or pay on time for the next loan is very high



Analyze Data on Application Data

Distribution of Clients' Age

- Very Young: 0 - 20 years old
- Young: 20-30 years old
- Middle Age: 30-50 years old
- Old: 50-60 years old
- Very old: >60 years old

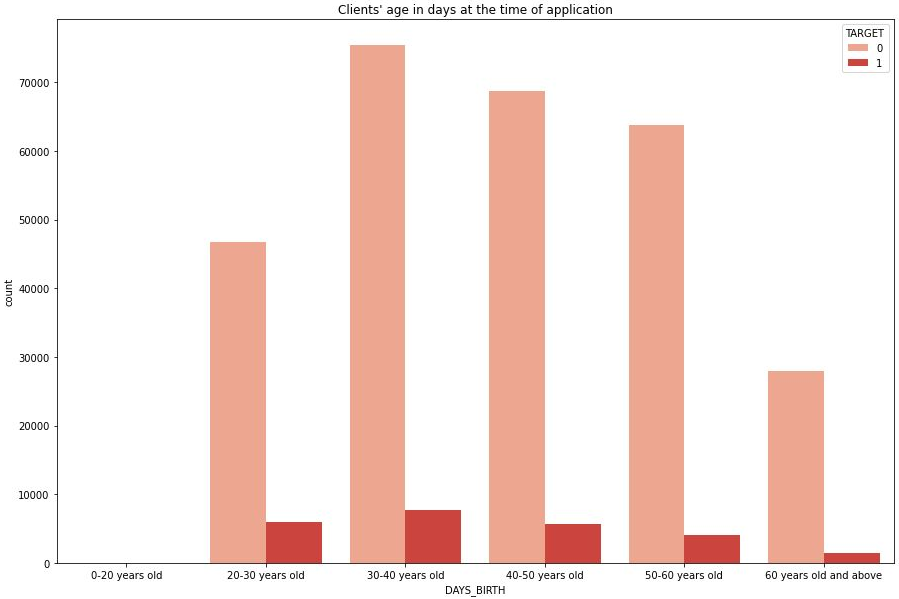
Insight:

After dividing customer age into various ranges, we can see:

- More than 50% applicants are in Middle age.
- 20.1% applicants are in Old age.
- 17.17% applicants are in Young age
- 9.55% applicants are in Very Old age
- Very few applicants are in Very Young age (<0.001%).

=> Overall:

- **The majority of loan applicants are middle-aged, followed by the Old, Young, Very Old and finally very young. 30-40 years old is the group with the most number of loan applications.**
- **20-30 years old group has fewer loan applications than the group of 40-50 years old or 50-60 years old, but the percentage of customers repaid on time is higher and almost equal to the 30-40 years old group.**



Analyze Data on Application Data

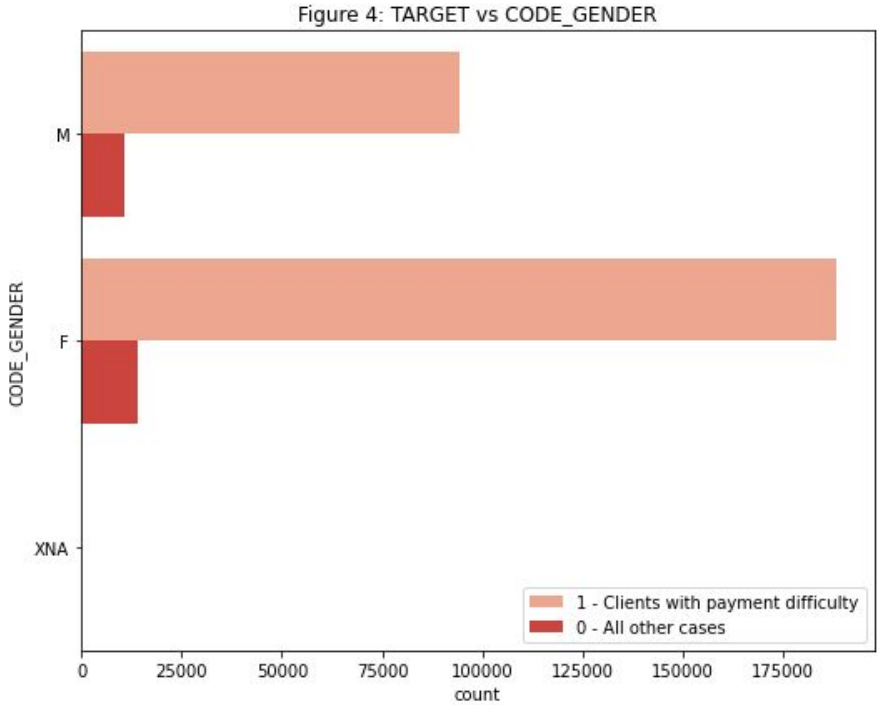
Distribution of Gender

Insight:

- Percentage of Female and Male Client are 65.83% and 34.16%
- XNA answer is very small, only have 0.001301%, so does not affect the difference of the remaining 2 answers too much. We can ignore this

=> Overall:

- **Total female clients is higher than the number of male clients (almost double).**
- **However, the on-time repayment rate of men is almost same as women.**



Analyze Data on Application Data

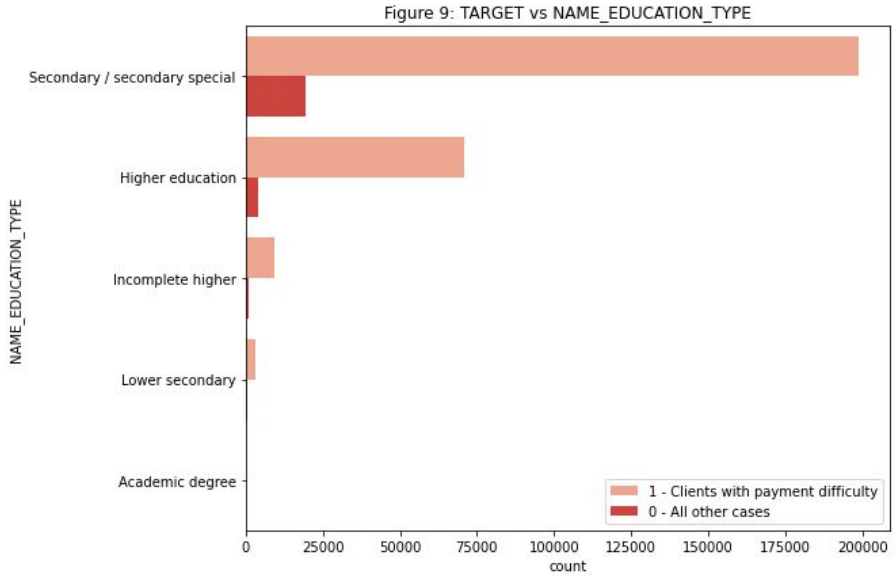
Distribution of Education

Insight:

- More than 71% clients have Secondary/secondary special.
- 24.34% clients have Higher education, followed by Incomplete higher and Lower Secondary
- Very small clients have Academic degree (only 0.05%)

=> Overall:

- **Secondary/secondary special category have the the largest rate of not paying loan on time.**
- **Clients with higher degree seem to be better able to repay their debts on time than others.**



Analyze Data on Application Data

Distribution of Income

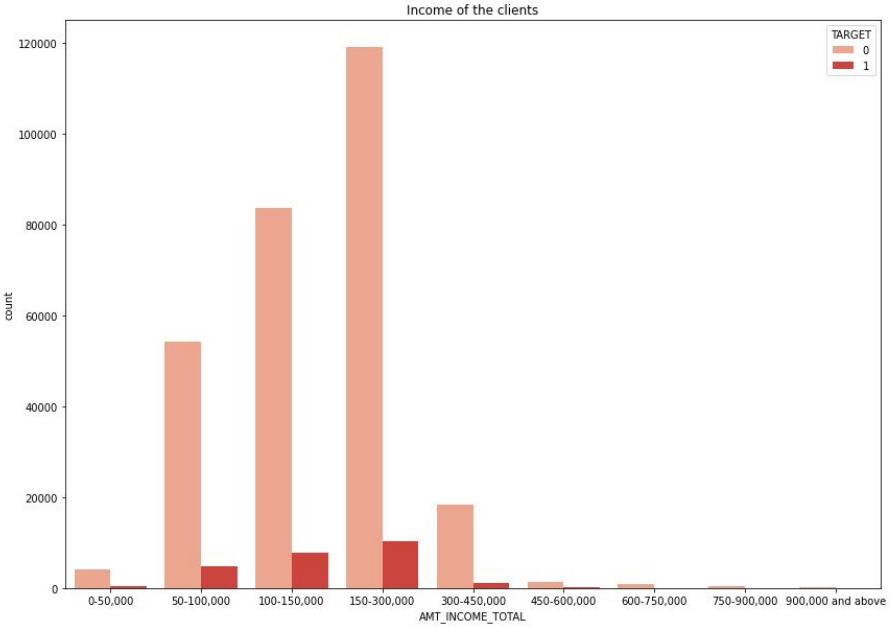
Insight: After dividing customer income into various ranges, we can see:

- More than 78% applicants from Low Income Group.
- 20.7% applicants from Very Low Income Group.
- 6.87% applicants from Medium Income Group
- 0.42% applicants from High Income Group
- Only 0.08% applicants from Very High Income Group.

=> Overall:

- **Loan applications are mainly in the Low income group. However, it is also the group that has most late repayment clients. In details, the applicants have income amount in the range of 150-300,000 has the most demand to apply for loans.**
- **High and Very High income groups have the least need for loans, but 100% of them are in the group of late repayment.**
- **The very low income group, although not having much number of loan applications, has a better proportion of on time repayment than others.**

- Very low: 0-100.000
- Low: 100 - 300.000
- Medium: 300 - 600.000
- High: 600 - 900.000
- Very high: >900.000



Analyze Data on Application Data

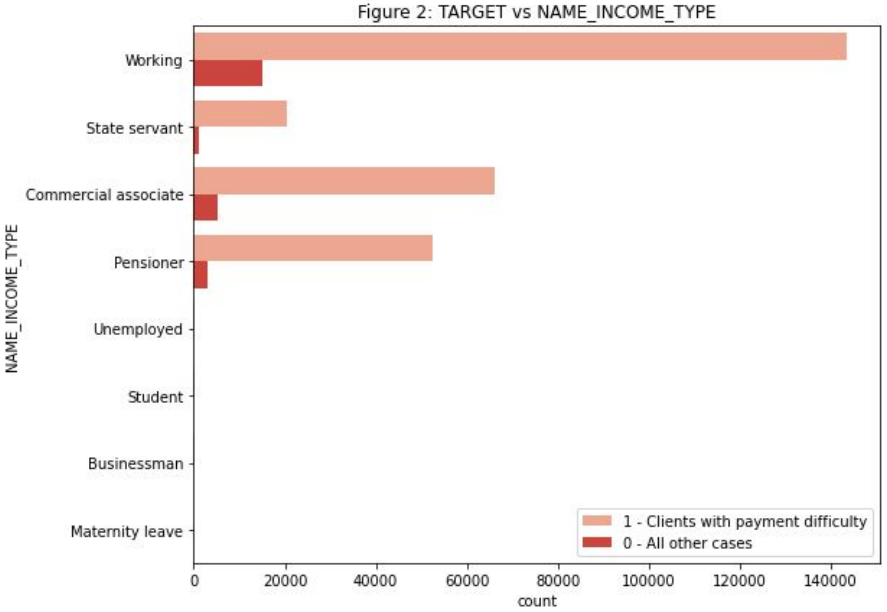
Distribution of Income type

Insight:

- 51.63% applications are people who have income type as Working
- 23.3% applicants have income type as Commercial associate
- Followed by Pensioner with 18% and State servant with 7%
- Unemployed, Student, Businessman, Maternity leave are very small (overall less than 0.008%)

=> Overall:

- **Working income type has has the largest percentage of non-refundable, followed by Commercial associate.**
- **State servant has a relatively safe rate of on-time payment compared to Pensioner.**
- **Unemployed, Student, Businessman, Maternity leave have less in numbers of applications**



Analyze Data on Application Data

Distribution of Clients' Working Time Range

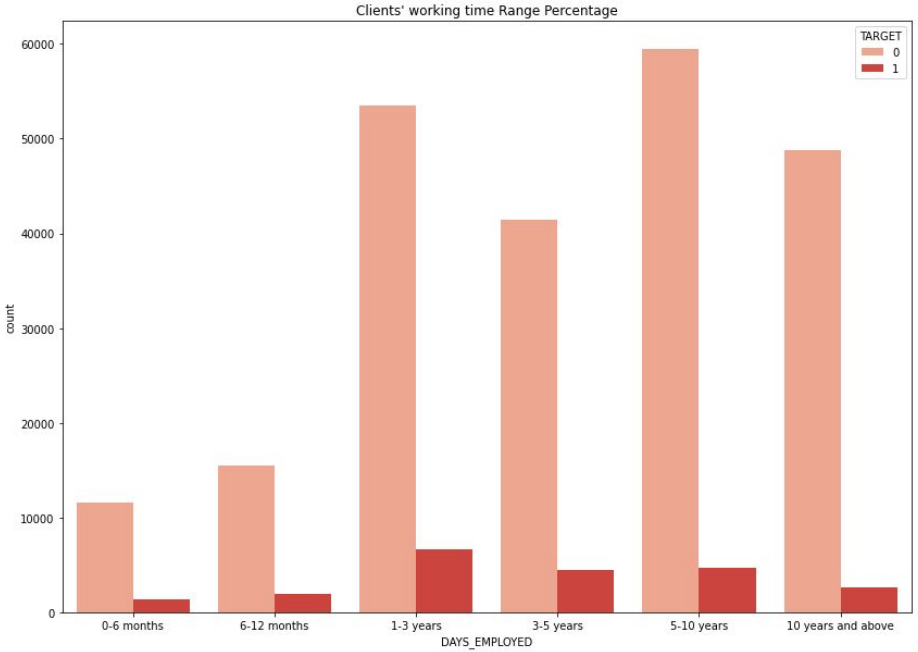
Insight: After dividing customer working time into various ranges, we can see:

- 25.5% applicants have long-time workings.
- 30.76% applicants have short-time workings.
- 20.42% applicants have very long-time workings.
- 18.21% applicants have medium time workings.
- 5.14% applicants have very short-time workings.

=> Overall:

- **Most loan applicants have working experience.**
- **Customers who pay their debts on time in the group of applicants with 1-3 years of working are the highest, although this is not the group that submits the most loan applications.**
- **Ratios in total clients who repay on time in the group of applicants with 3-5 years of working are equal to the group of 5-10 years of working, and higher than the the group of over 10 years of working, although the group of over 5 years of working takes up more**

- Very short-time: 0-6 months
- Short-time: 6 months - 3 years
- Medium: 3-5 years
- Long-time: 5-10 years
- Very long-time: >10 years



Analyze Data on Application Data

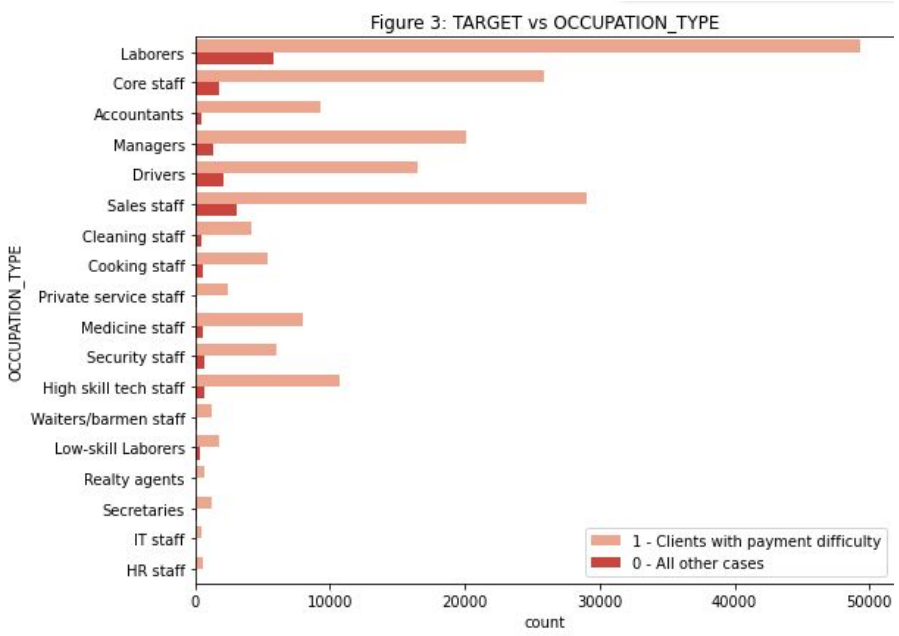
Distribution of Clients Occupation

Insight:

- 26.14% loans applications are from Laborers
- Sales staff, Core staff, Managers account for about 10 - 15%/per type
- Drivers, High skill tech staff, Accountants, Medicine staff, Security staff, Cooking staff, Cleaning staff, Private service staff account for about 1 - 8%/per type
- Low-skill Laborers, Waiters/barmen staff, Secretaries, Realty agents, HR staff, IT staff, each type less than 1%

=> Overall:

- **Jobs that require qualifications have a higher percentage of people who do not pay their debts on time than jobs that do not require too many qualifications.**



Analyze Data on Application Data

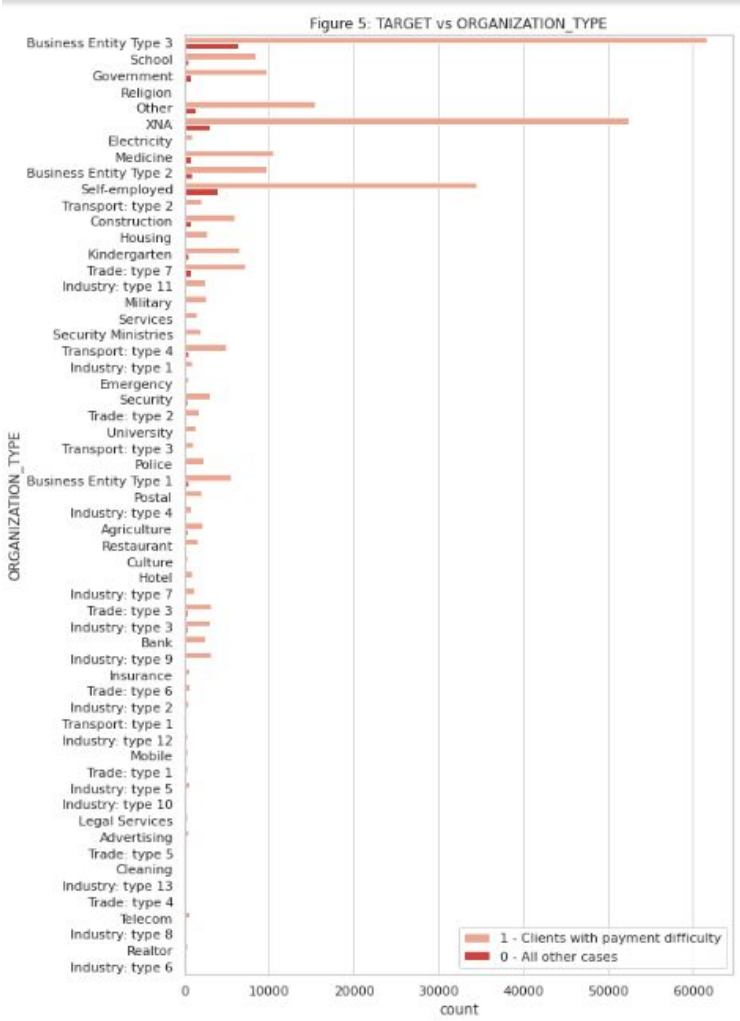
Distribution of Clients Organization Type

Insight:

- Business Entity Type 3 has highest percent of loans applications, followed XNA, Self-employment.
- Other, Medicine, Business Entity Type 2, Government, School, Trade: type 7, Kindergarten, Construction, Business Entity Type 1, Transport: type 4, Trade: type 3, Industry: type 9, Industry: type 3, Security have 1% - 5%/per category number of loan applications

=> Overall:

- **Those who are Business Entity (especially type 3), startup or not providing answer (XNA) high probability of default**



Analyze Data on Application Data

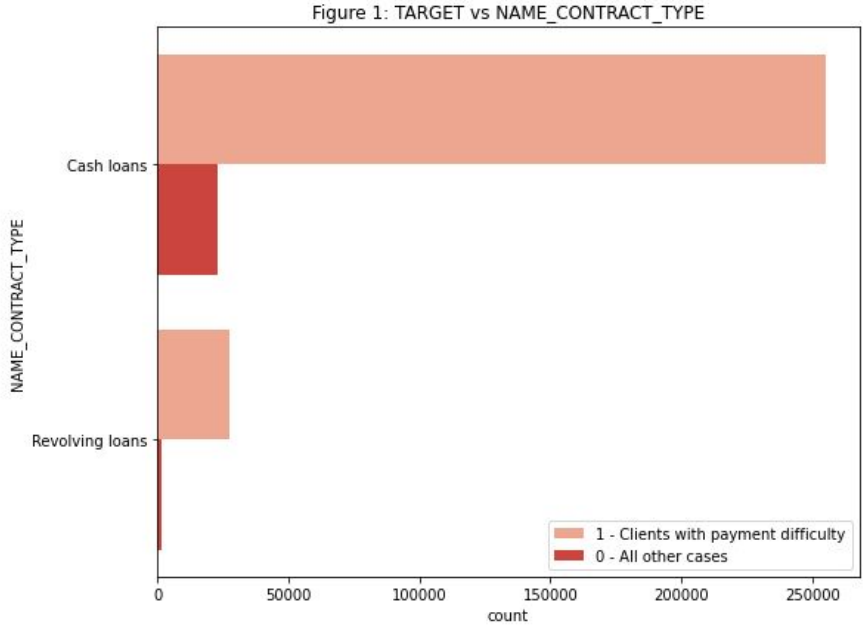
Distribution of Contract Type

Insight:

- Over 90% of the loan is Cash loans
- Revolving loans is only about 9.5%

=> Overall:

- **Cash loans have a high risk of non-repayment or default.**
- **Revolving loans account for a small portion of total loan applications, however, if looking at the difference between non on-time and on-time payments of revolving loans, it is much smaller than cash loans.**



Analyze Data on Application Data

Distribution of Credit Amount of the Loan

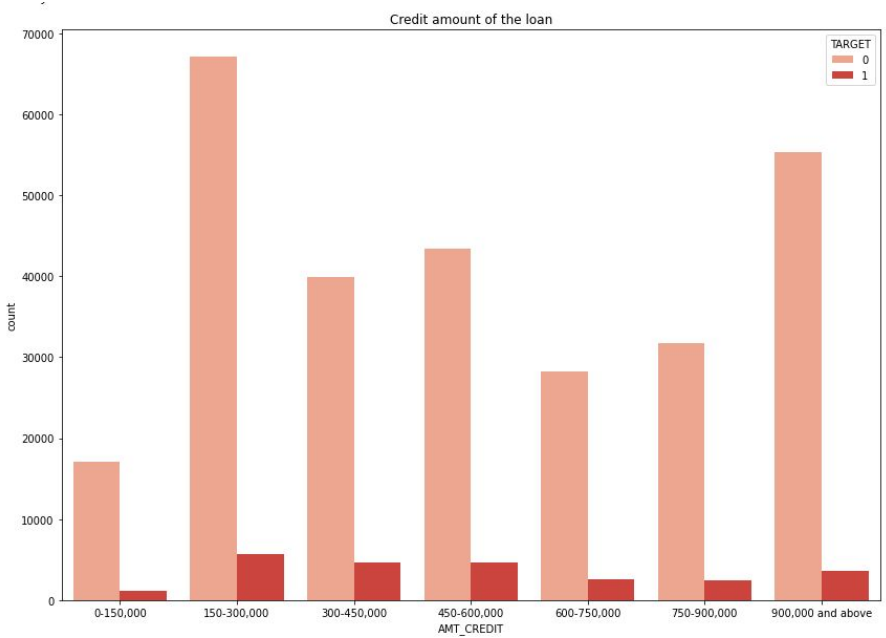
Insight: After dividing Credit Amount of the Loan into various ranges, we can see:

- 23.7% applicants applies low credit amount of loan.
- 30.12% applicants applies medium credit amount of loan.
- 21.14% applicants applies high credit amount of loan.
- 19,16% applicants applies very high credit amount of loan.

=> Overall:

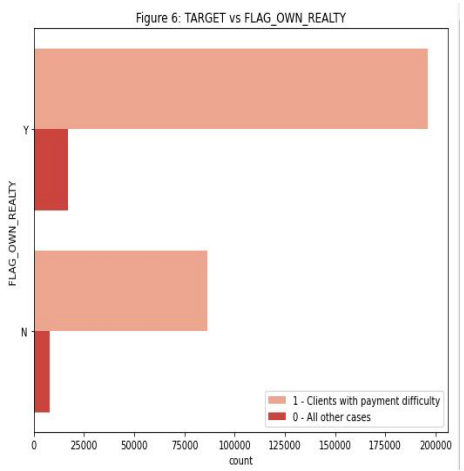
- **Borrowers from Very Low credit are the group with the lowest rate of non-payment on time**
- **The risk of not repaying on time or defaulting is high in the group of loans from 150-300.000 (Low) and over 900.000 (Very High).**
- **Medium credit group has a fairly high amount of on-time repayment, not much difference between the price ranges in this group.**

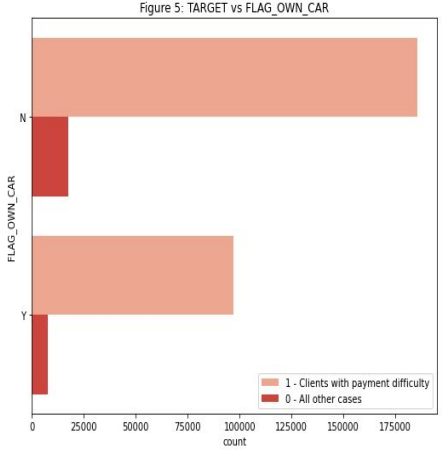
- Very low: 0-150.000
- Low: 150 - 300.000
- Medium: 300 - 600.000
- High: 600 - 900.000
- Very high: >900.000

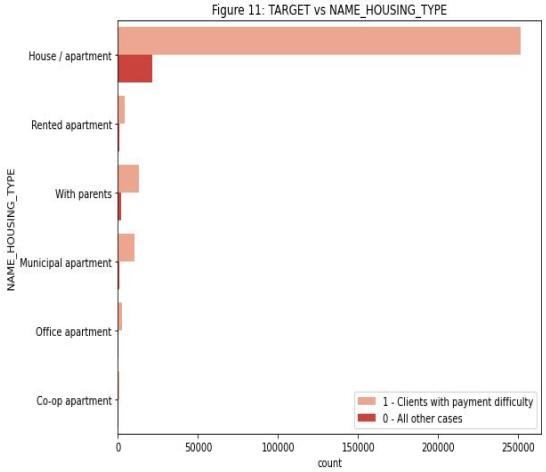


Analyze Data on Application Data

Distribution of Client assets







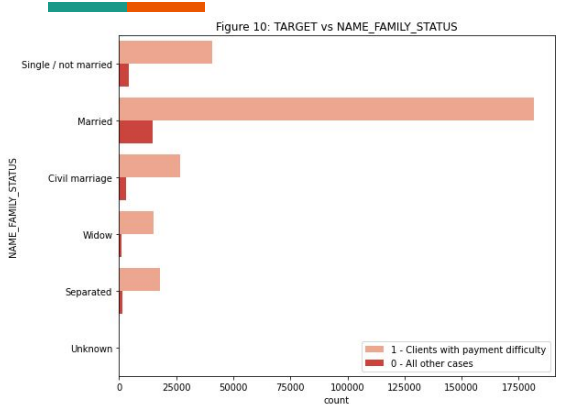
Insight:

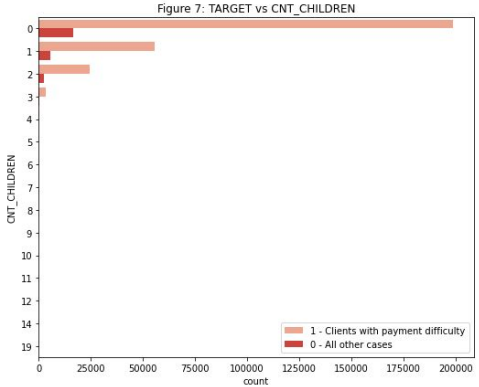
- Majority of people have their own realty or car, live in house/apartment apply for the loan
- People live in co-op apartments have no default rate

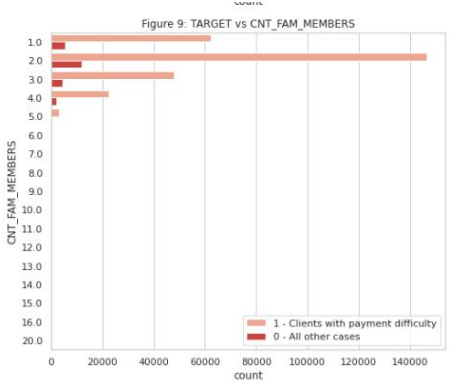
=> Overall: People with property and vehicles are the most applying for loans. However, the group of having assets have a higher risk of default. Whereas, the group of no car or no realty has better performance on returning loans.

Analyze Data on Application Data

Distribution of Family Information







Insight:

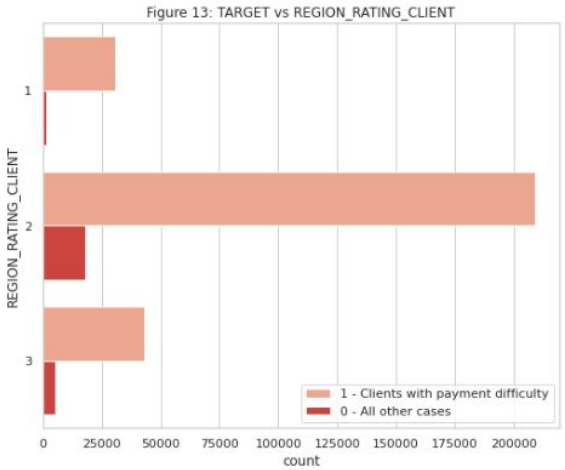
- People who are married have highest number of not returning loan on time.
- Very less clients who have more than 3 children or 5 family apply for loans

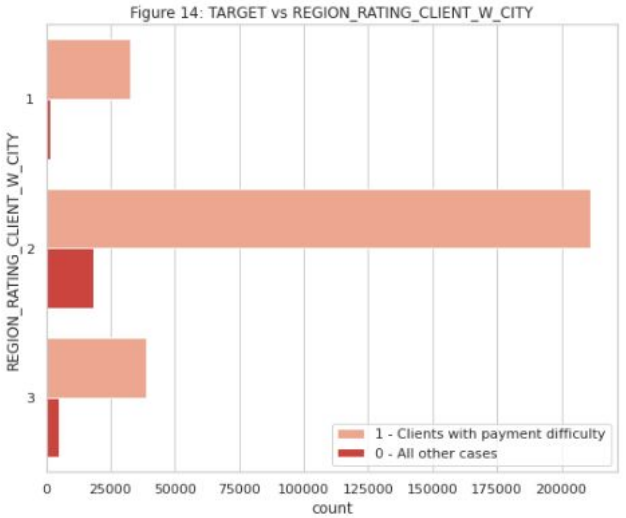
=> Overall:

- **Most loan applicants are getting married and have no children.**
- **Clients with 2 family member, no children have the risk of defaulting**

Analyze Data on Application Data

Distribution of Area Information





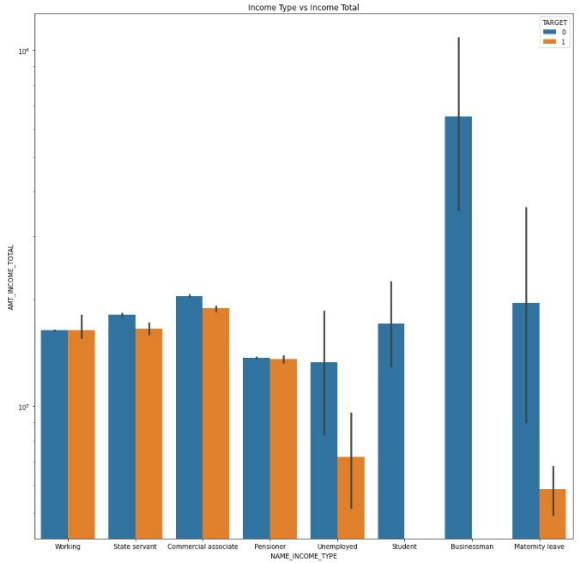
Insight: About 74% the applicants live in 2

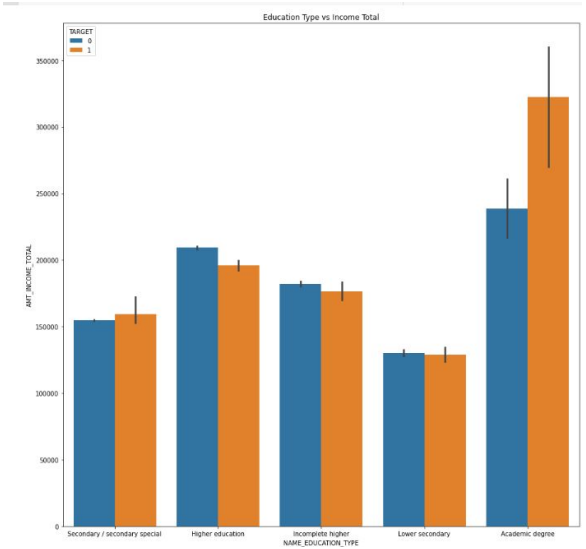
=> Overall:

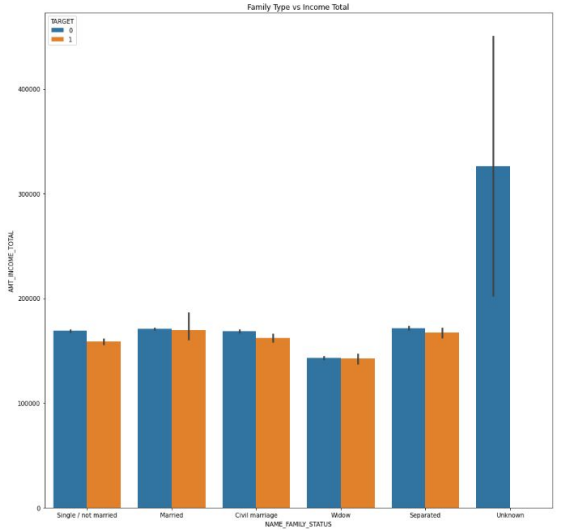
- **Region Rating 2 has the highest default rate**
- **Region Rating 1 has the lowest default rate**

Bivariate Analysis on Application Data

Distribution of income type/education type/family type vs Income total





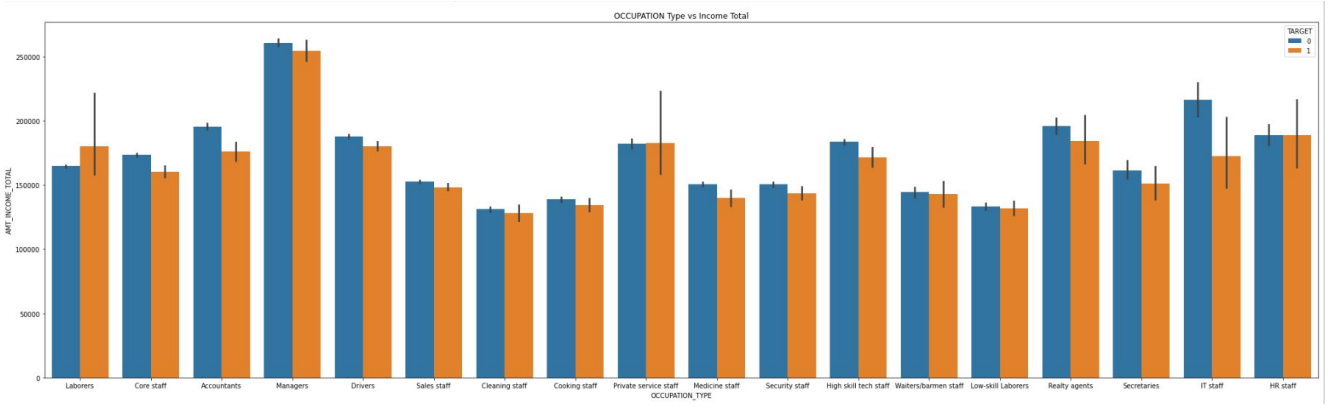


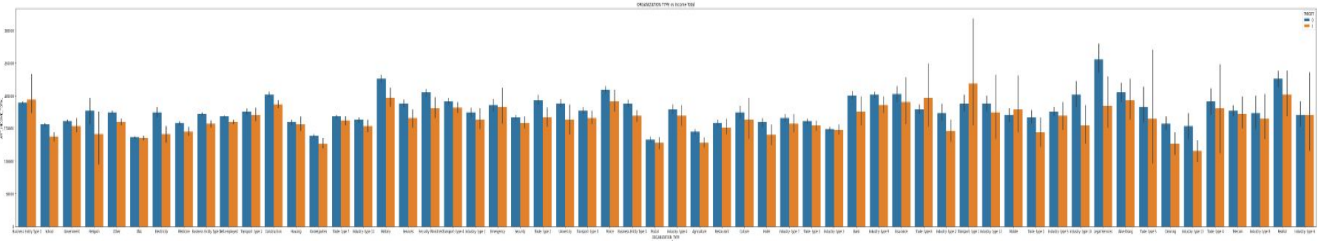
Insight:

- Income Type: Businessman has highest income but 100% of them are in non returning payment.
- Education Type: Academic Degree has highest income, but people that pays on time has a higher income
- Family status: Unknown category has highest rate in non on time payment. However they are also in the highest income group.

Bivariate Analysis on Application Data

Distribution of occupation type/ organize type vs Income total



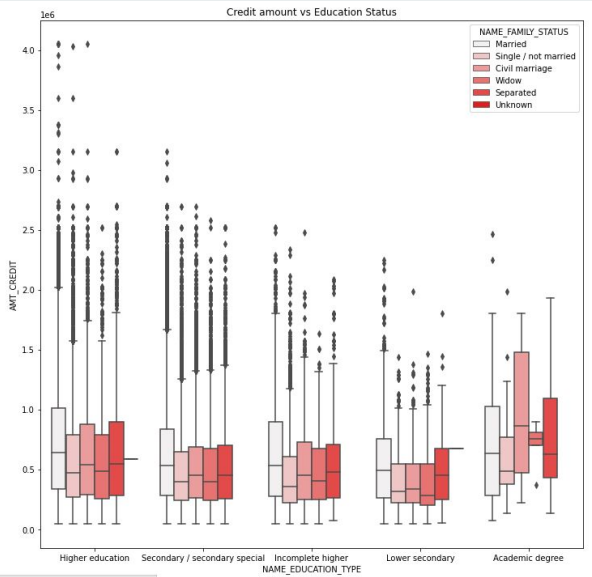


Insight:

- Occupation Type: Manager has highest income but mainly belong to the group of not paying debts on time
- Organization: Legal service has highest income. For Business Type 3 and Transport Type 1 groups, the higher the income, the better the ability to repay on time.

Bivariate Analysis on Application Data

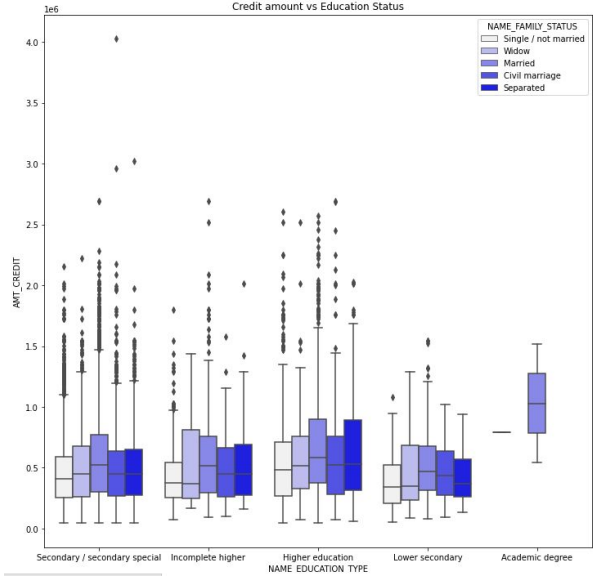
Distribution of Clients education type vs credit amount vs family status - Target 0



Insight:

- In Academic degree group, Civil marriage type applies higher credits than others
- In Higher education group, Married type applies higher credits than others and has the most outliers

Distribution of Clients education type vs credit amount vs family status - Target 1



Insight:

- In Academic degree/ Secondary/ Secondary special/ Higher education group, Civil marriage applies higher credits than others
- In Higher education group, Civil marriage has the most outliers

Summary:

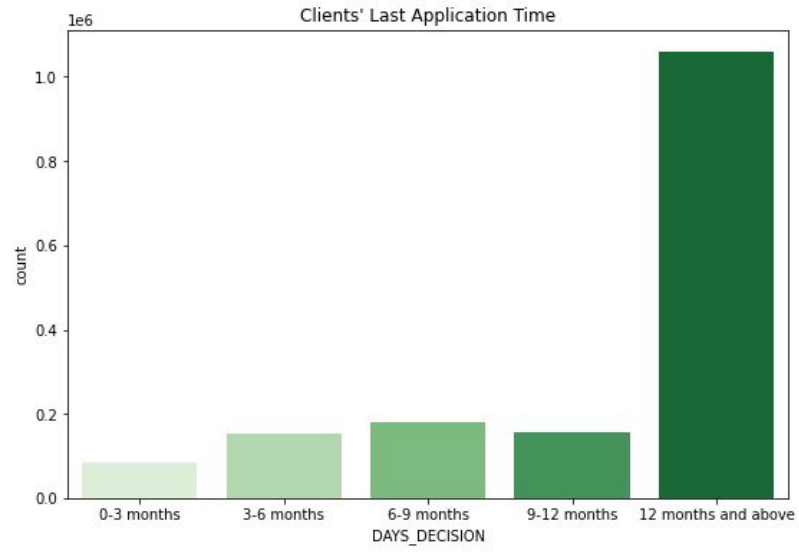
- GENERAL INFORMATION:
  - AGE: Customers in the Young age group (20-30 years old) are better able to repay their loans on time
  - GENDER: Female clients have higher risk of not returning their loans.
  - EDUCATION: Clients with high degree seem to be better able to repay their debts on time than others.
  - FAMILY: People with more family members (especially only have 2 members) have a higher risk of default. People who are married or do not provide family status have highest number of not returning loan on time.
  - ASSETS: The group of having assets have a higher risk of default. Whereas, the group of no car or no realty has better performance on returning loans.
  - REGION RATING: Considering clients who are living in Region Rate 2 because they have the highest default rate
- INCOME:
  - Need to have a lot of considerations when approving loans for low, high and very high income groups.
  - People with high positions seem to be safe for loan approval

Summary:

- WORK EXPERIENCES:
  - Applicants with 1-5 years of working have better repayment ability than the rest of the groups.
  - Jobs that require qualifications have a higher percentage of people who do not pay their debts on time than jobs that do not require too many qualifications.
  - Applicants from Transport Type 1 have ability to repay on time.
  - Those who do not provide the type of organization or have their own business have a high default rate
- LOAN TYPE:
  - Revolving loans have better on-time repayment rates than cash loans.
  - The group of Credit amount of the loan from Low credit and Very high credit have high rate of non ontime payment and default. Group borrowing less (lower than 150,000) will be less risky, Borrowers from 300-600.000 (Medium) have a higher ability to repay on time

Analyze Data on Previous Application

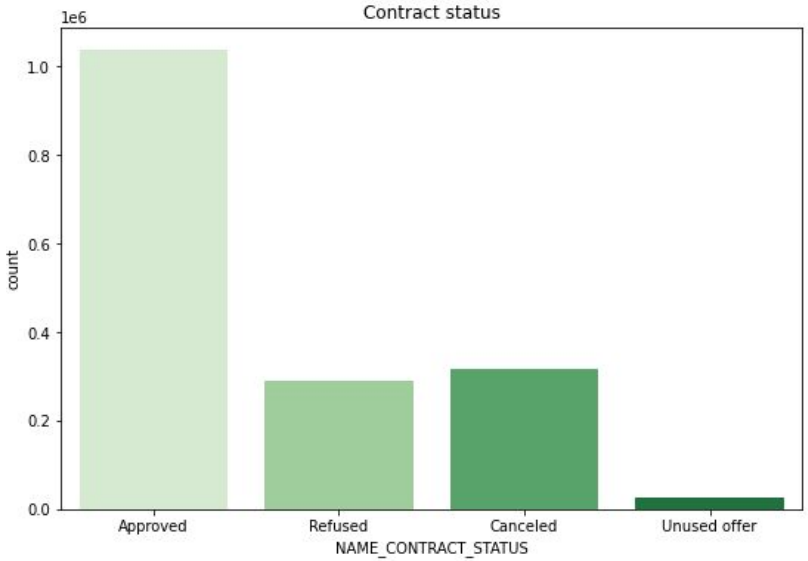
Distribution of Previous Application Time



Insight:

- Most of them come back to apply loans after 12 months, followed by 6-9 months.

Distribution of Previous Contract Status



Insight:

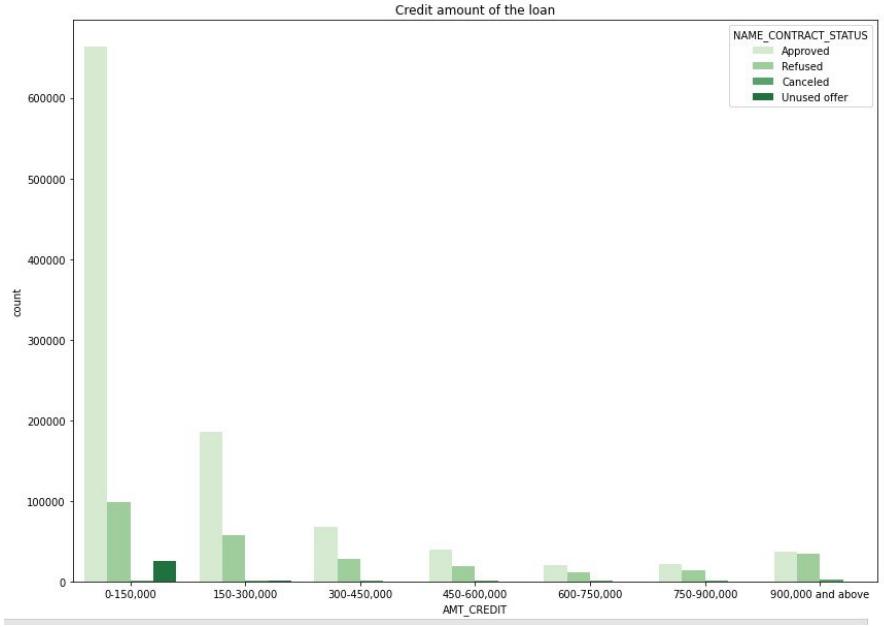
- Most of applications were approved
- Cancel rate was too high

Analyze Data on Previous Application

Distribution of Previous Application Credit Amount of loan

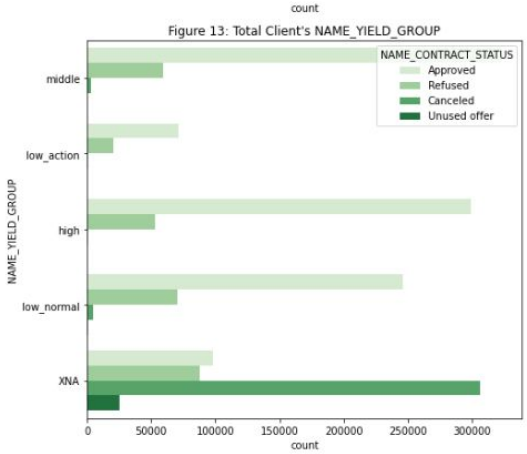
Insight:

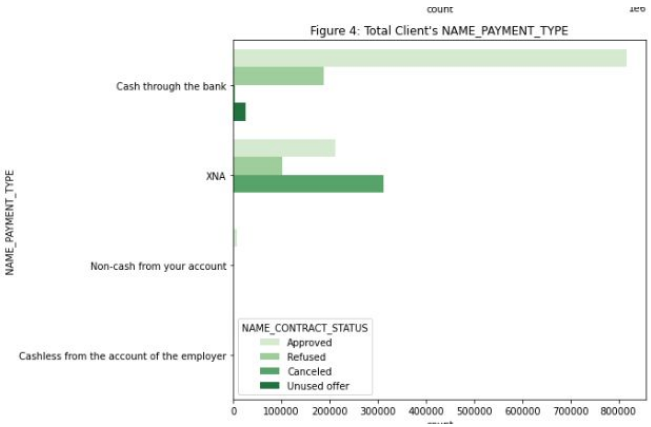
- More than 59% applicants applied very low credits (less than 150.000) for previous application and most of them were approved.
- The reject rate fell mainly on very low, low and very high credit.
- The cancellation rate is very low
- Very low credits group has the highest unused offer rate

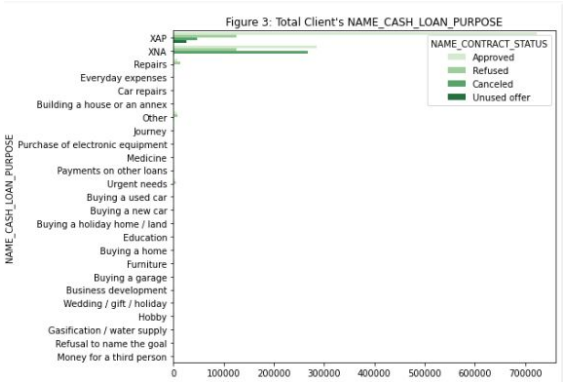


Analyze Data on Previous Application

Distribution of Previous Contract Information





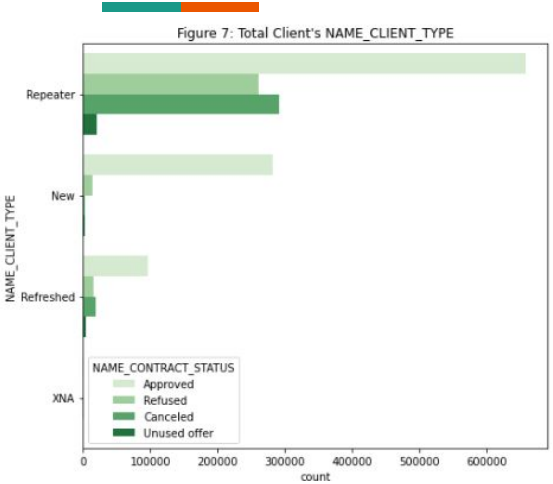


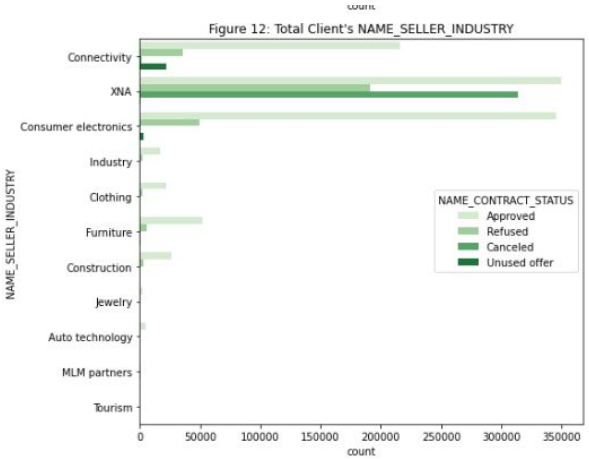
Insight:

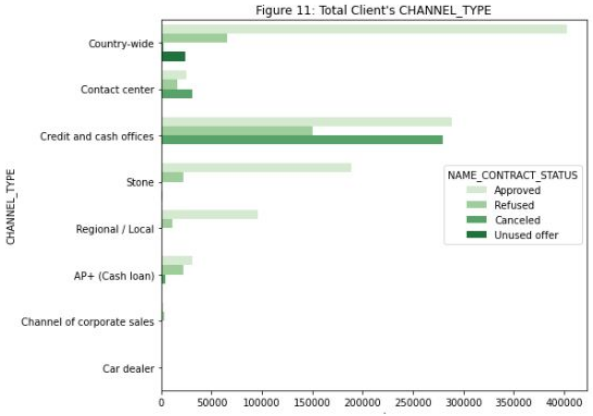
- The group that didn't provide yield group answers is highly likely to be rejected and self-cancelled
- Middle and high interest rate group in previous application have a high chance of being approved
- The group that didn't provide payment type answer is highly cancelled. Cash through the bank payment got rejected a lot
- The group that provide XAP or XNA answer on Cash loan purpose question got a high refused rate, but XAP group still had the highest approvement

Analyze Data on Previous Application

Distribution of Previous Customer Information







Insight:

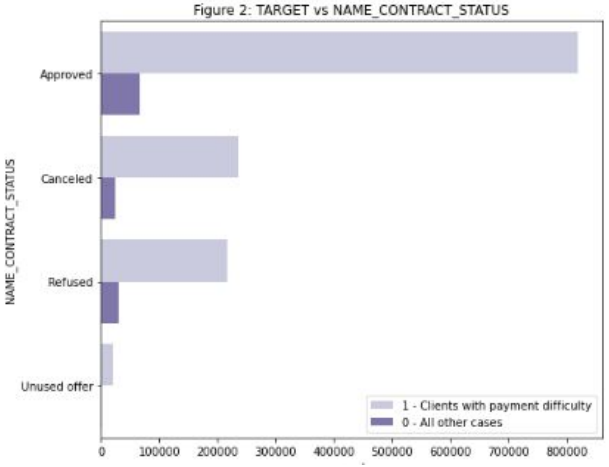
- Mostly old customers, this group also has the highest cancellation rate
- New customer groups are less likely to be rejected
- The group that didn't provide industry is highly likely to be rejected and self-cancelled. Consumer electronic group had a high chance of being approved
- Clients from Country-wide mostly approved, whereas credits and cash officers were often cancelled and refused.

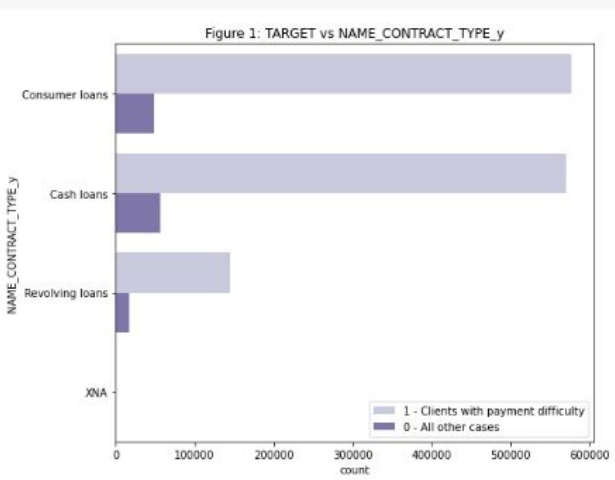
Summary:

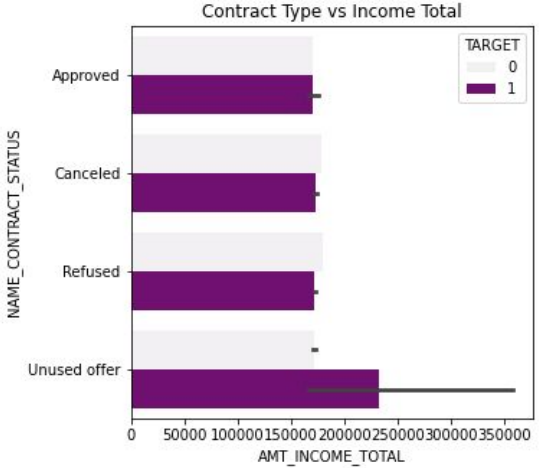
- For small borrowers (less than 150.000) and Repeat customers, a lower interest rate or better pricing should be provided to limit the cancellation rate
- Customers who do not provide all the answers are more likely to be denied or change their mind and cancel their request.

Multivariate analysis - Analyze Data on Application Data and Previous Application

Distribution of Contract Status





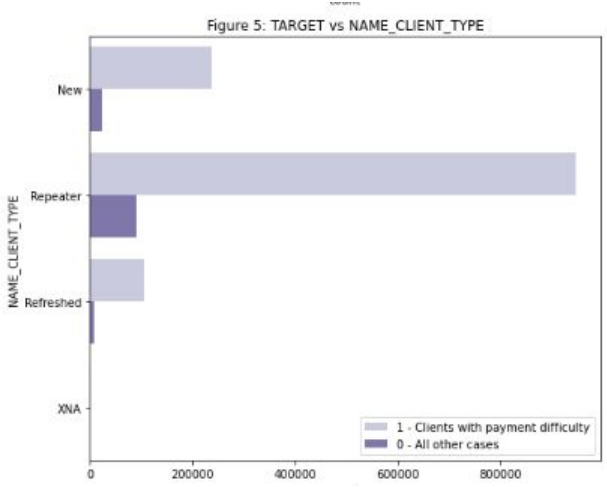


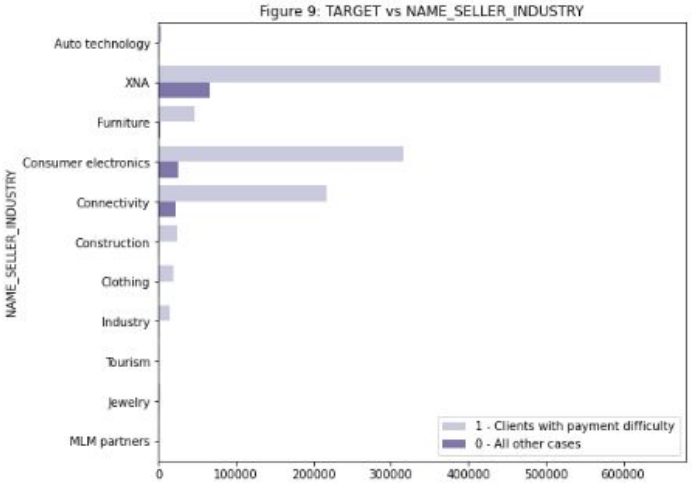
Insight:

- Out of 62% of customers approved on the previous loan, most of the customer group does not pay the loan on time.
- The group of customers who borrow money in the form of loan revolving is still the least risky group of customers
- Unused offer group on previous time has the highest income and highest rate of on-time payment
- 17.35% people who were rejected the application in the previous loan application came back to apply again

Multivariate analysis - Analyze Data on Application Data and Previous Application

Distribution of Customer Information



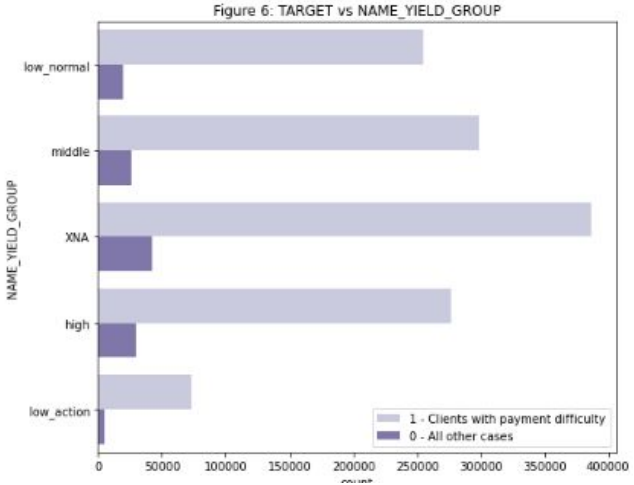


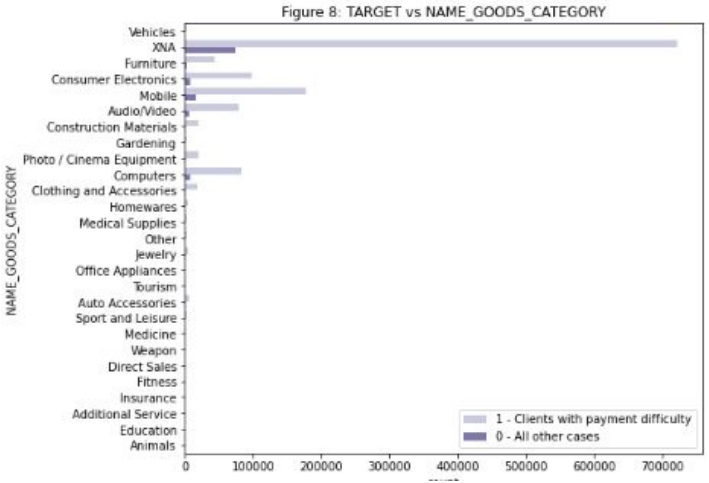
Insight:

- Most of the old customer group does not pay the loan on time.
- The group that didn't provide industry is highly likely to not returning.
- Consumer electronic group and Connectivity group has most approvement on previous but it also has the highest non repaying rate after XNA group

Multivariate analysis - Analyze Data on Application Data and Previous Application

Distribution of Loan Information



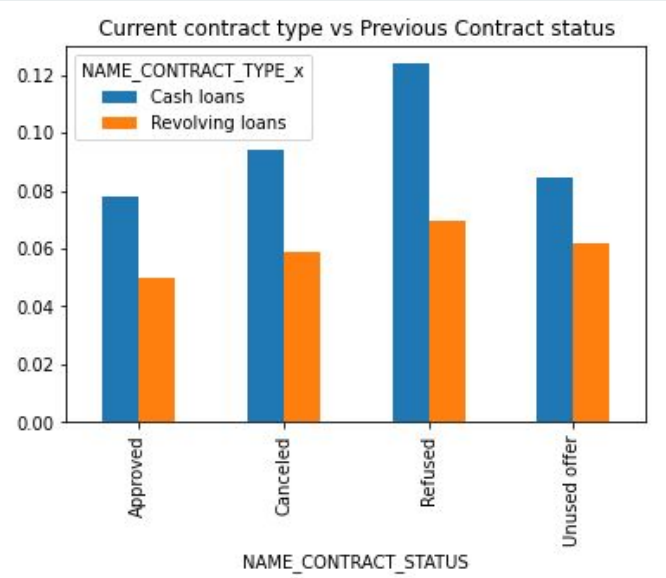


Insight:

- The group that didn't provide yield group and goods category is highly likely to not returning.
- High group and middle group has most approvement on previous but it also has the highest non repaying rate after XNA group

Multivariate analysis - Analyze Data on Application Data and Previous Application

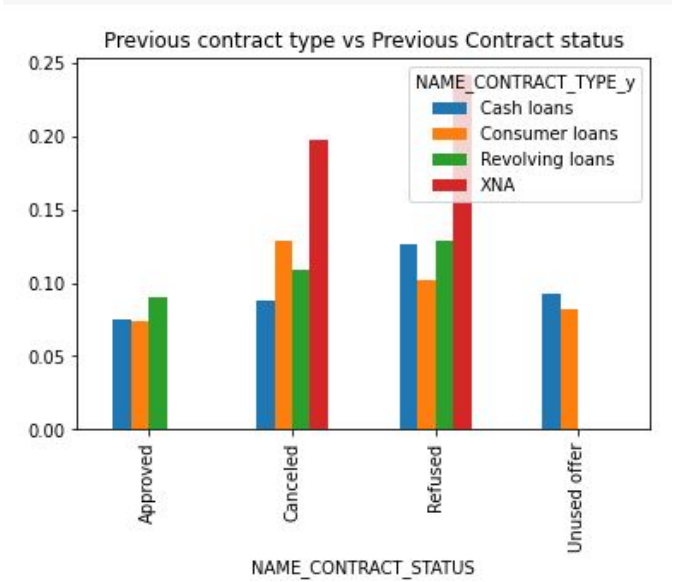
Distribution of Current contract type vs Previous Contract status



Insight:

- Most of cash loans applications belong to the group that was rejected or cancelled in the previous loan.
- The ratio of group borrowing in the form of revolving loans is not much different

Distribution of Previous contract type vs Previous Contract status

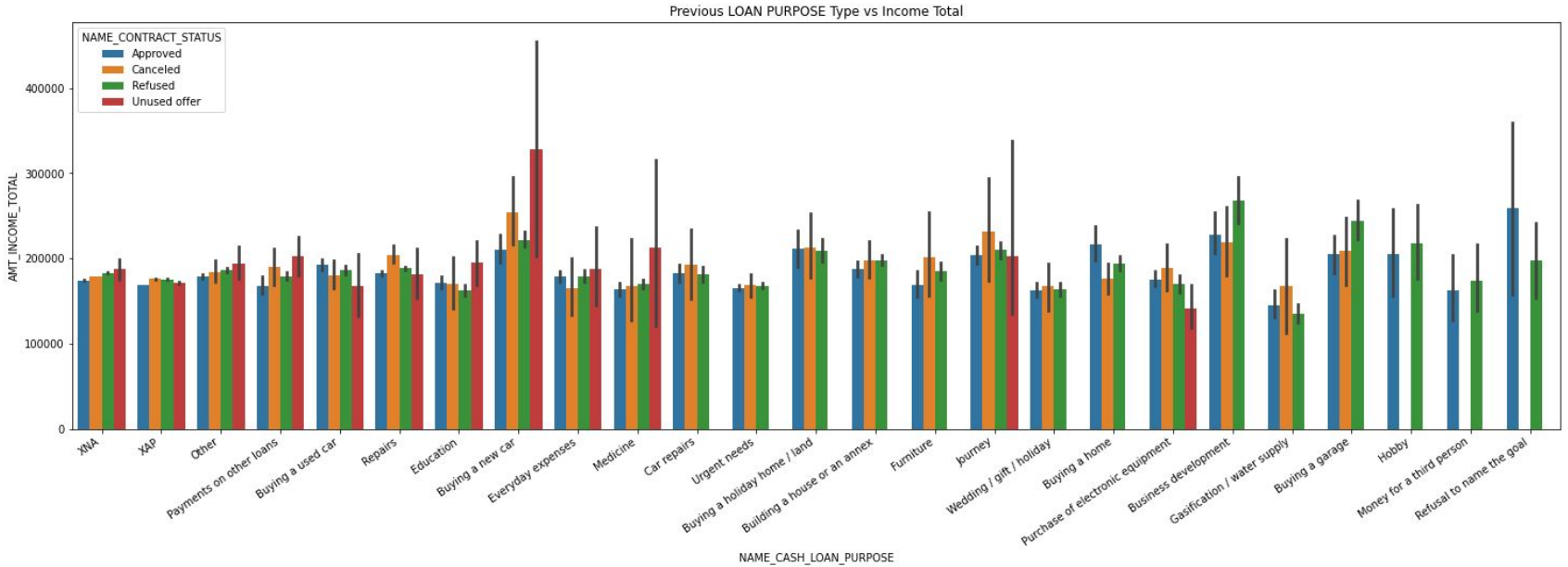


Insight:

- People who provide XNA answer have high risk of denial and cancelation.
- Revolving loans group has no risk of stopping the processing contract.

Multivariate analysis - Analyze Data on Application Data and Previous Application

Distribution of Previous Cash Loan Purpose Type vs Income Total - TARGET 0

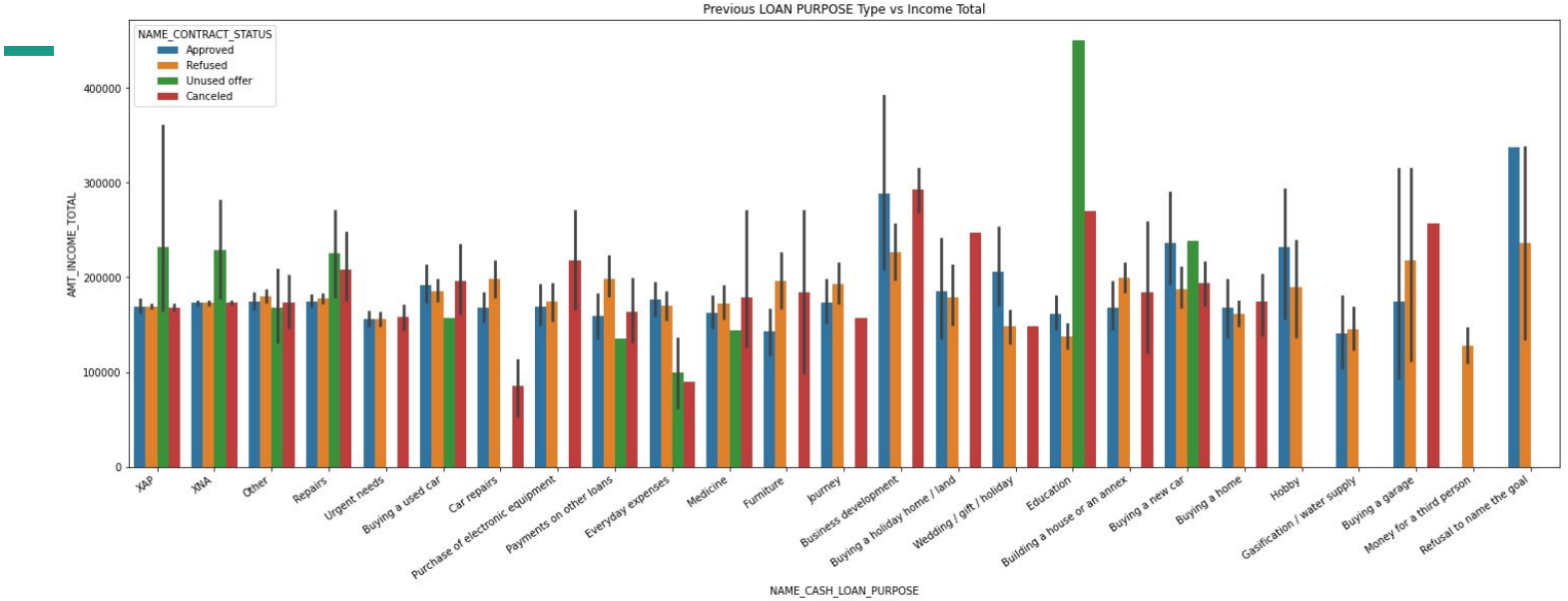


Insight:

- Those who have a credit loan to buy a car, have the highest income. However, these people canceled during process.
- Following is the Business development group with the second highest level of income, but these salary are quite low and they were rejected by the company. The group that does not provide purpose is approved by the company also has quite low income

Multivariate analysis - Analyze Data on Application Data and Previous Application

Distribution of Previous Cash Loan Purpose Type vs Income Total - TARGET 1

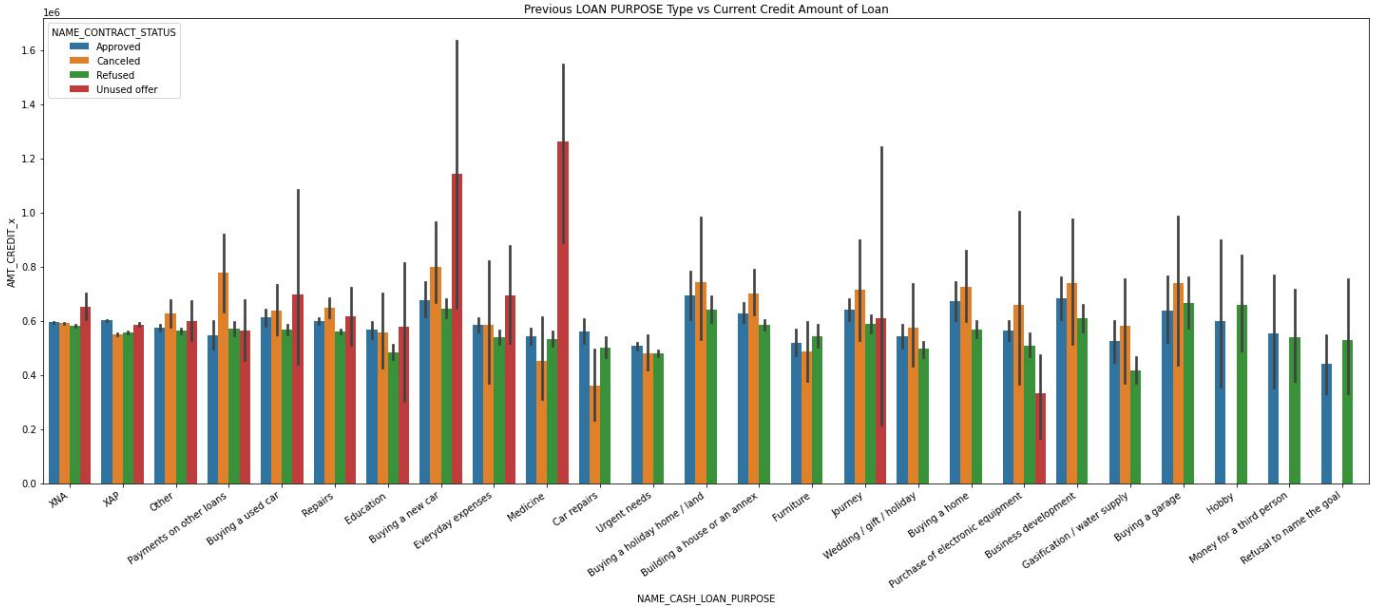


Insight:

- High rate of Unused Offer of education group had highest income. Whereas, Refusal to name the goal still had medium income and rate of passing on previous application.

Multivariate analysis - Analyze Data on Application Data and Previous Application

Distribution of Previous LOAN PURPOSE Type vs Current Credit Amount of Loan - TARGET 0

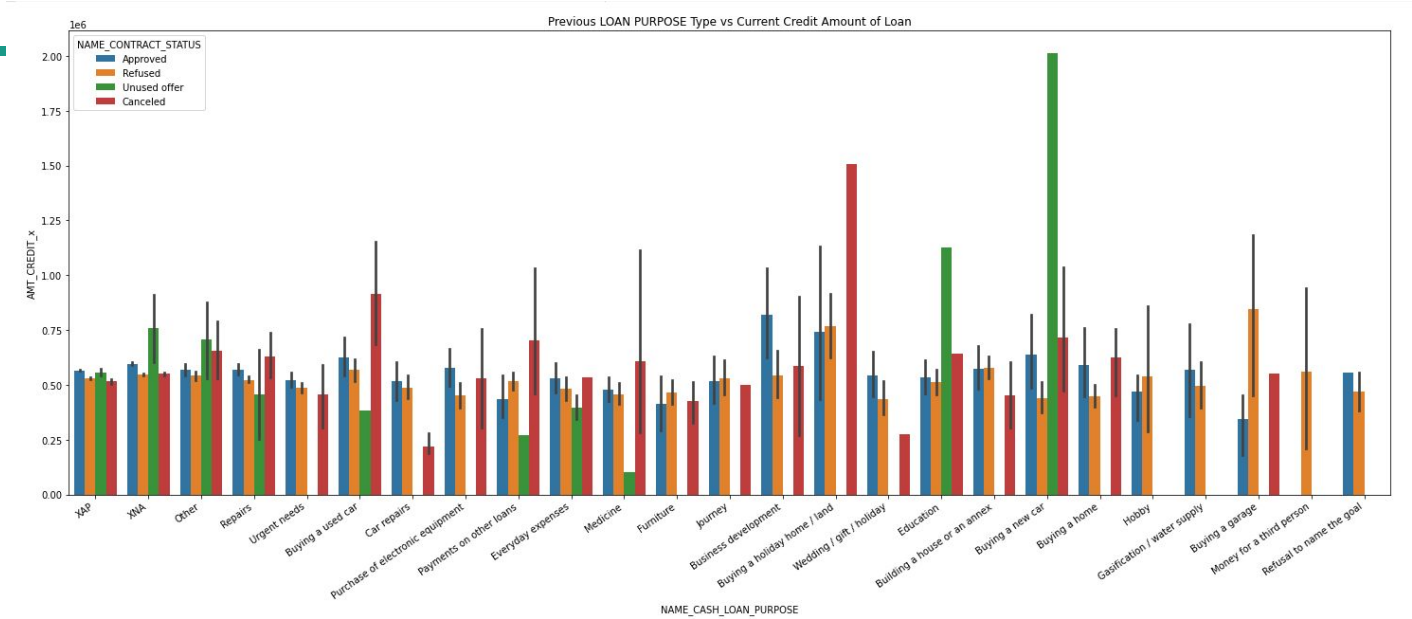


Insight:

- Medicine and Buying a new car purpose applies highest credits but cancelled previous application
- High credits rate of Previous Rejected Applications was in Buying a garage and Hobby options

Multivariate analysis - Analyze Data on Application Data and Previous Application

Distribution of Previous LOAN PURPOSE Type vs Current Credit Amount of Loan - TARGET 1

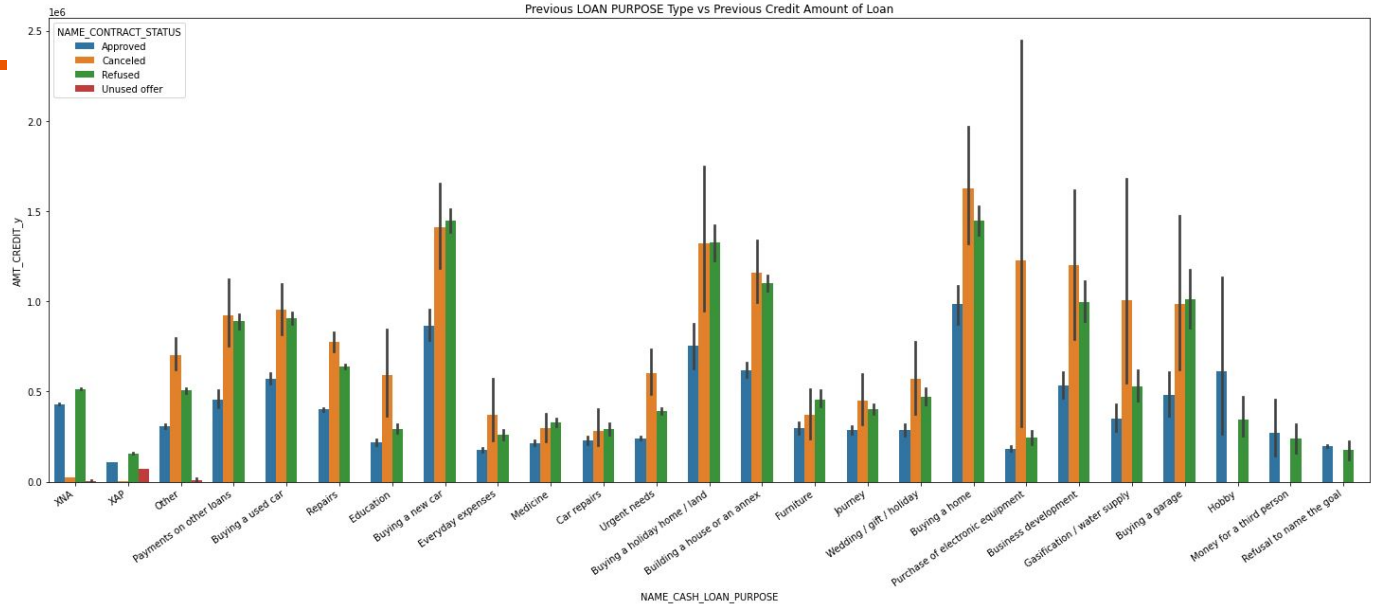


Insight:

- The group of customers applies very high credit to buy a new car but canceled the contract during the completion process.
- The group of customers applies high credit to buy a holiday home/land but cancelled the application

Multivariate analysis - Analyze Data on Application Data and Previous Application

Distribution of Previous LOAN PURPOSE Type vs Previous Credit Amount of Loan - TARGET 0

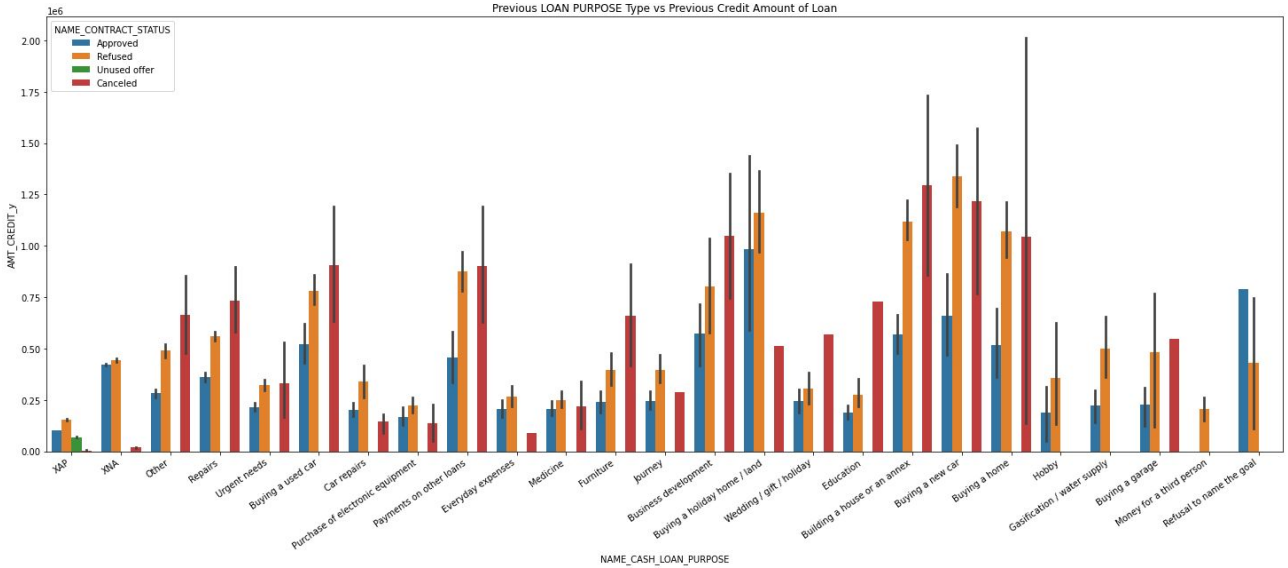


Insight:

- The unused offer group appears very rarely in this chart.
- Same as previous slide, the group of customers applies highest credits to buy a home but cancelled the application during approval
- Most of the applications that borrow high or very high loans focus on the material procurement group (such as estate, vehicles,...)

Multivariate analysis - Analyze Data on Application Data and Previous Application

Distribution of Previous LOAN PURPOSE Type vs Previous Credit Amount of Loan - TARGET 1

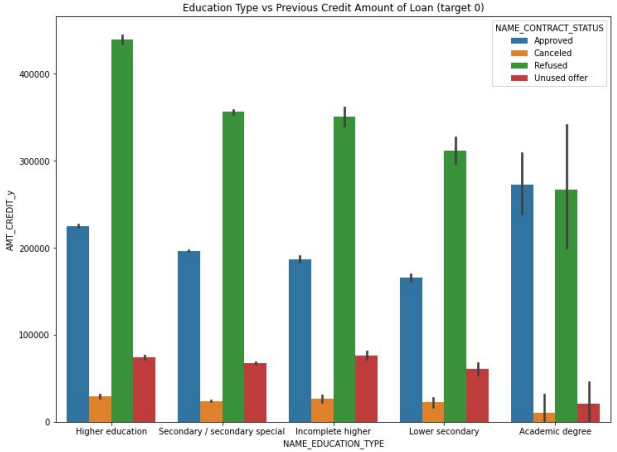


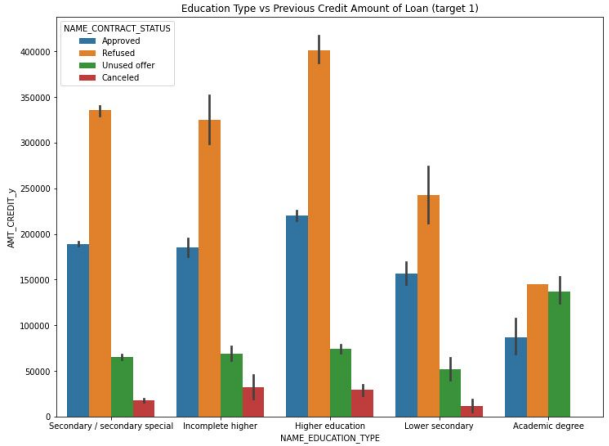
Insight:

- The unused offer group only appears at XAP answer
- The group of customers applies highest credits to buy a car but cancelled the application during approval
- Those who are approved are usually low-credits applicants

Multivariate analysis - Analyze Data on Application Data and Previous Application

Distribution of Previous Education Type vs Previous Credit Amount of Loan



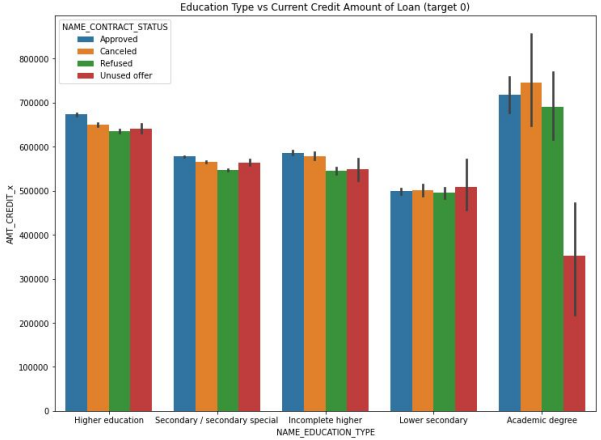


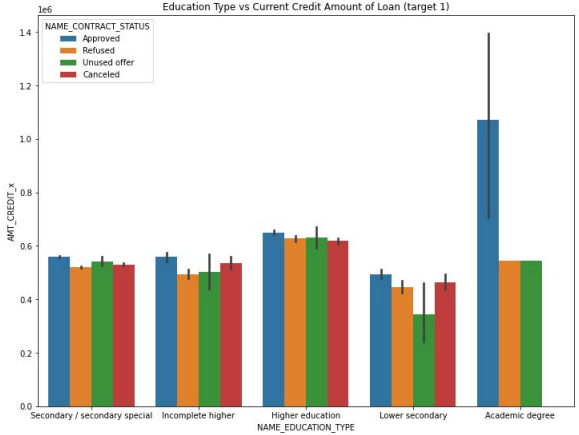
Insight:

- For target 0: customers with higher education applied for medium credits if comparing to total (the highest credits if only counted in the group 0) were rejected. Only people who have academic degree can apply successfully for high credit
- Higher education type in Group 1 is same as Group 0

Multivariate analysis - Analyze Data on Application Data and Previous Application

Distribution of Previous Education Type vs Current Credit Amount of Loan



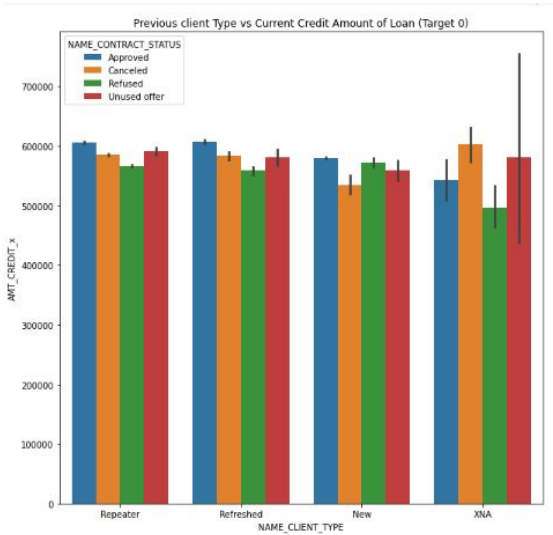


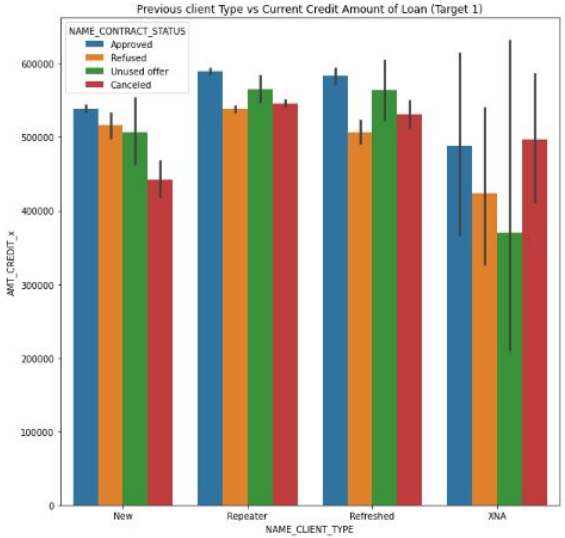
Insight:

- Both 1 and 0, only people who have academic degree with successful previous application continue to apply for higher credits
- Those who have canceled their contracts in the previous, apply for a much higher amount of credit at this time

Multivariate analysis - Analyze Data on Application Data and Previous Application

Distribution of Client type vs Credit Amount





Insight:

- All client type groups have quite similar proportions.
- For Target 0: in refreshed customers, loan credit is approved at a medium price, followed by Repeater and New. Customer who listed in XNA approved lower credit but also in the medium range.
- For Target 1: for Repeated customers, loan credit is approved at a medium price, followed by refreshed (very close to Repeated) and New. Customer who listed in XNA approved lower credit but also in the medium range.
- Customers in group 1 will be approved for a loan with a higher credit level than group 0

Summary:

- Considering approving loan applications for old customers who have a history of not paying loans on time.
- Customers who do not willing to provide all the answers are more likely to delay repayment process.
- In general, the customers who cancel the contract (maybe because they change mind or the do not get a suitable rate as they expected) are the customers who apply loans to buy assets such as houses and cars, etc.
- Unused offer group on previous time has the highest income and highest rate of on-time payment
- Those who are able to pay their debts on time, are rarely in the group of Unused offers.
- Need to recheck those who had the ability to pay their debts on time and applied high credit amount of loan but are rejected on previous time to minimize the loss of potential customers.

Conclusion

- People who canceled their application in the previous loan application came back to apply again. Therefore, Lower interest rate or better pricing should be provided to limit the cancellation rate or dropped rate.
- Company should check the reasons for the rejection of the applicants who were denied on the previous loan but are able to repay the loan on time to decrease the business loss.
- Those that need to be considered or are at high risk of default should be rejected for loan or provide loan with higher interest rate or provide small credits loans to decrease the risk of defaulting.