RISK ANALYTICS IN BANKING AND FINANCIAL SERVICES (CREDIT EDA CASE STUDY)

- **Gaurav Gahlot**
- **Saurabh Swroop**

PROBLEM STATEMENT

To apply EDA in a real business scenario for developing a basic understanding of risk analytics in banking and financial services and minimizing the risk of losing money while lending to customers.

OUR APPROACH

We had 2 files which are as explained below:

1. 'application_data.csv' contains all the information of the client at the time of application.
The data is about whether a **client has payment difficulties.**

2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

We have done separate analysis for both the datasets we had. And then inferred important indicators for knowing how can we segregate the clients who might face difficulty in repaying back and can be a defaulter; hence refusing them the loan and those who are opposite of these and for them loan can be approved.

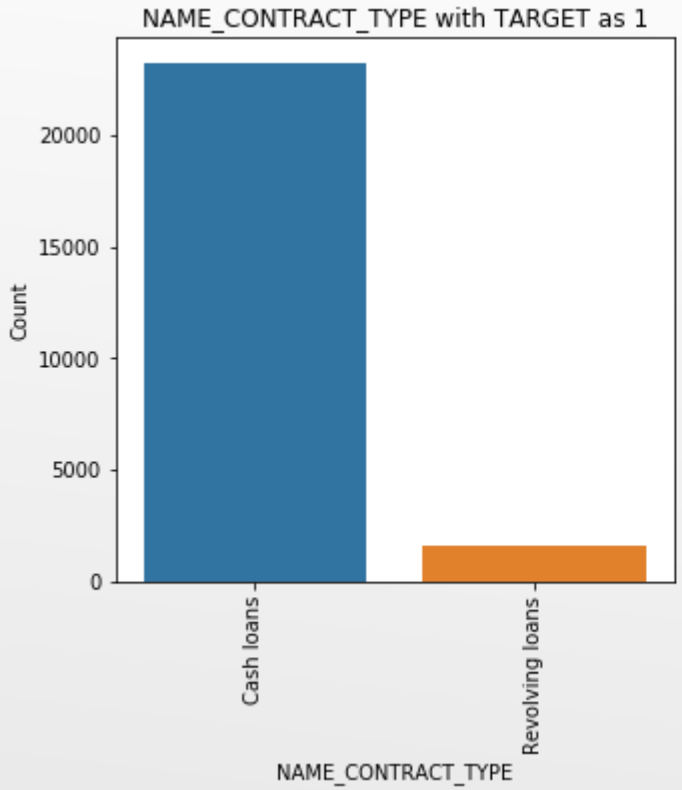
This is the basic approach we used while analysis:-

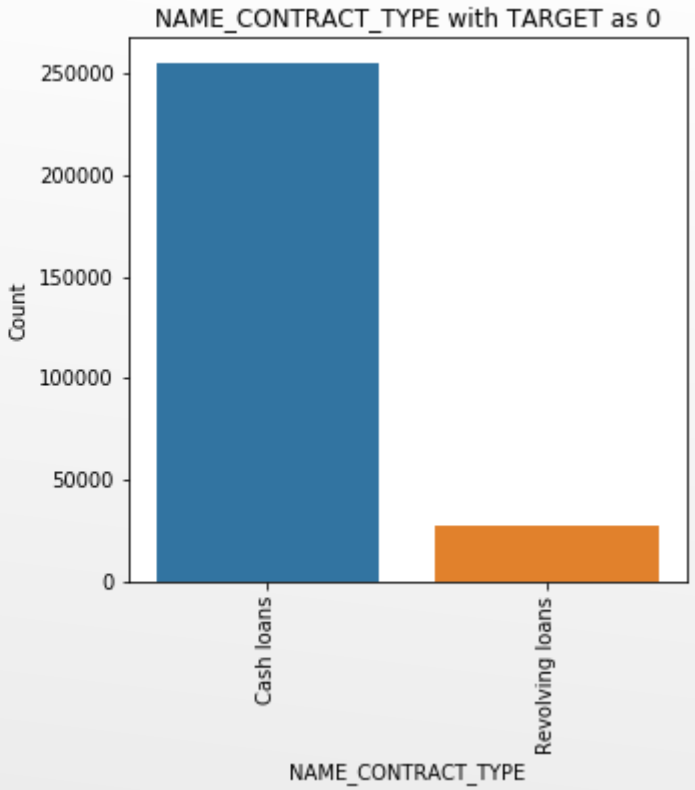
- Looked at the missing value percentages in the columns. This was very important as this helped in carrying out Attribute Reduction. We kept 50% as the threshold and so all those columns which had missing values more than or equal to 50% were dropped.
- Selected the columns which we found with good sense.
- We are suggested few methods for missing value imputations.
- Checked the datatypes for the attributes and changed if required.
- We looked at the numerical attributes for the outlier detection using box plots and also treated them.
- Binning of Continuous Variables.
- Based on the imbalance ratio, segregated the data into sub-datasets and then carried out the rest of the analysis (Univariate, Segmented univariate and Bivariate).

APPLICATION DATA EDA

- We will now 1st check the file which contains all the information of the client at the time of application.
- The data is about whether a **client has payment difficulties or not.**

PLOTS AND INFERENCES

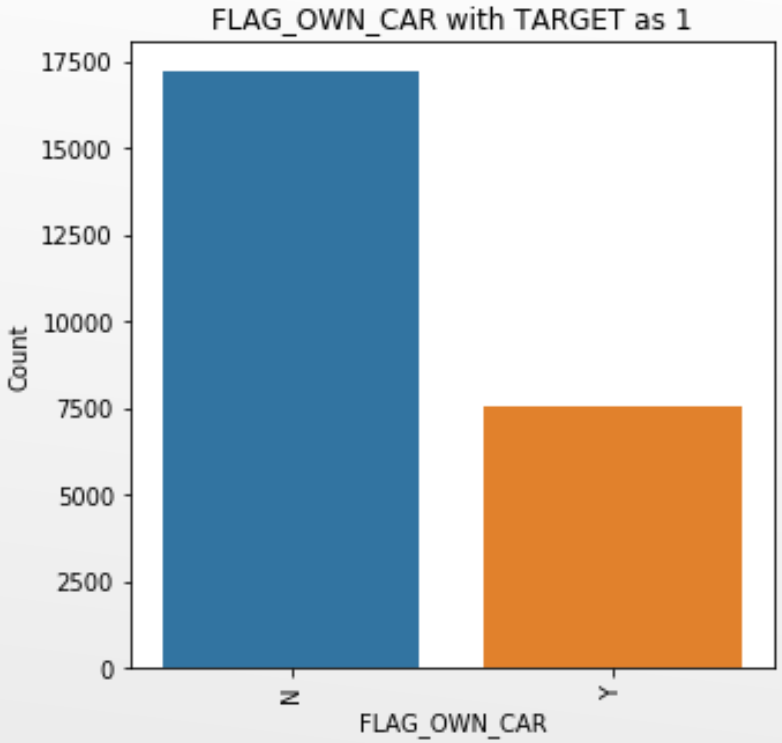


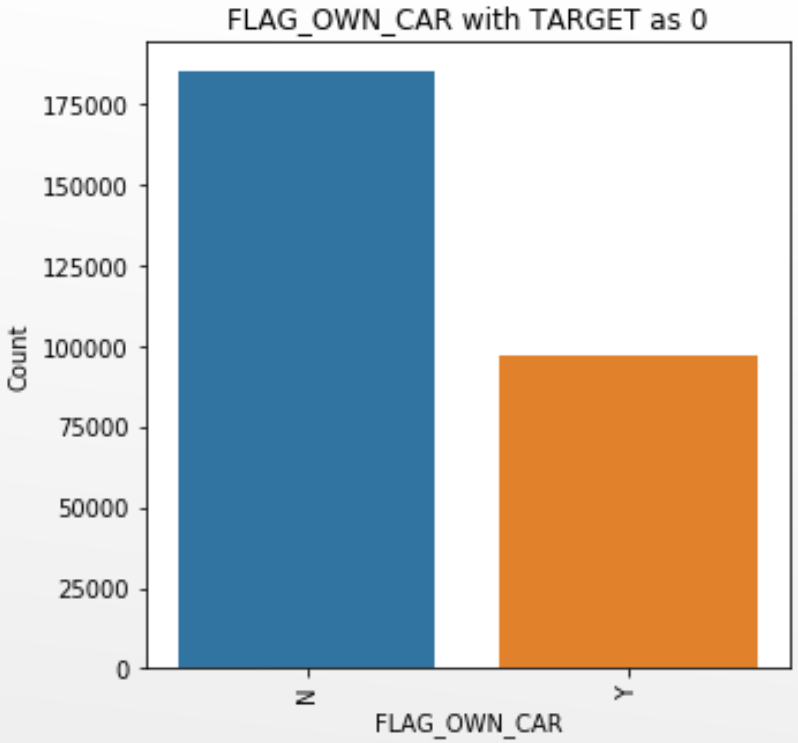


- In both the categories of TARGET, we see that there are more of the Cash loans in comparison to the Revolving loans (to say, it is more than 10 times).

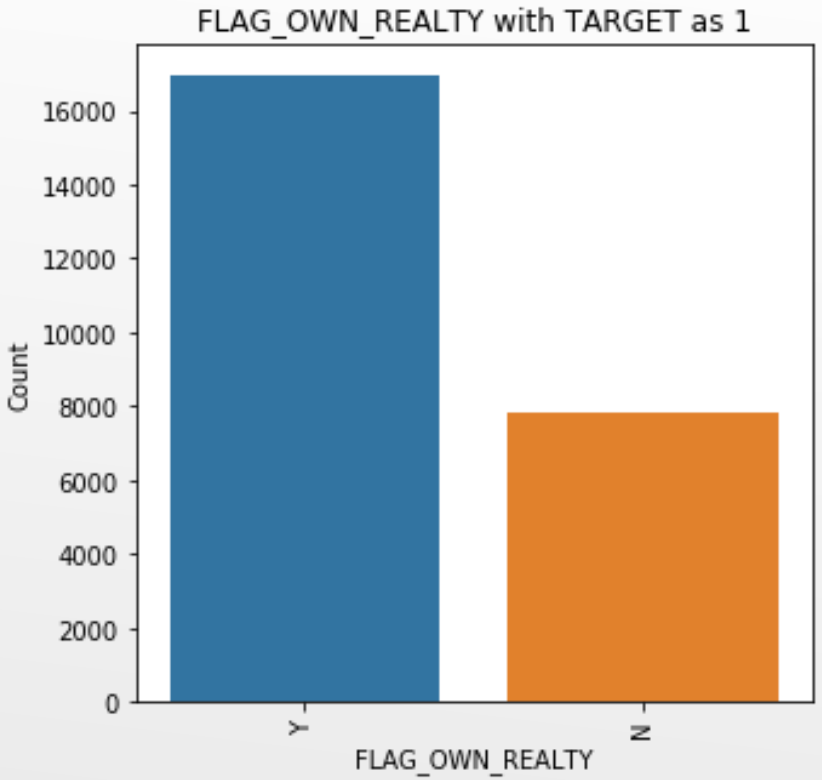
- Another thing we see is that be it any type of loan, there are 4 times less chances for a client to be a Cash loan holder; since the two bars in any of the graphs are with 4 times difference.

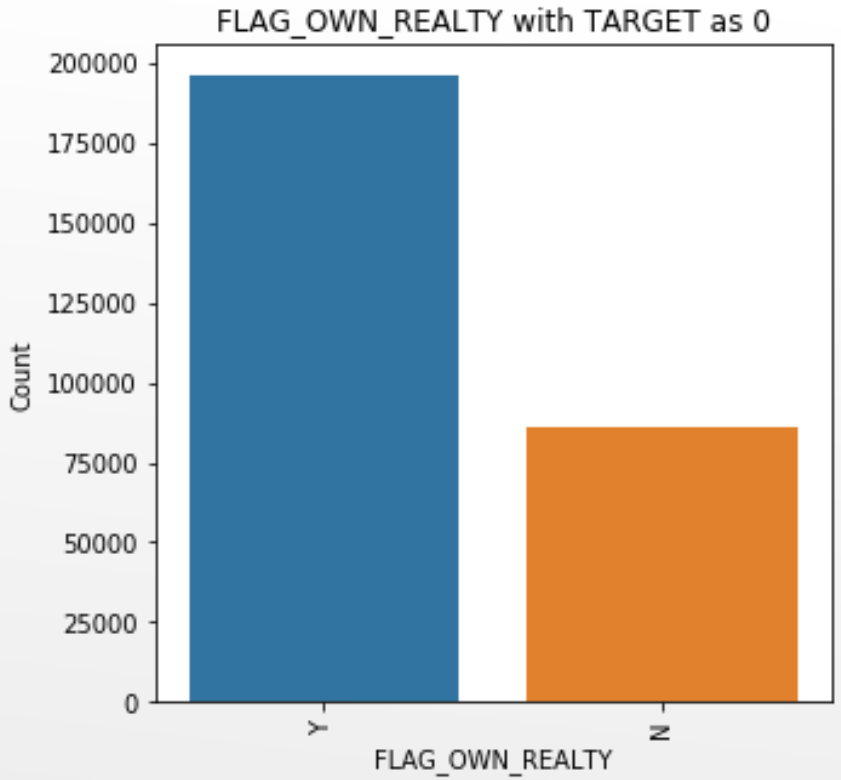
Females take more loans than the males.



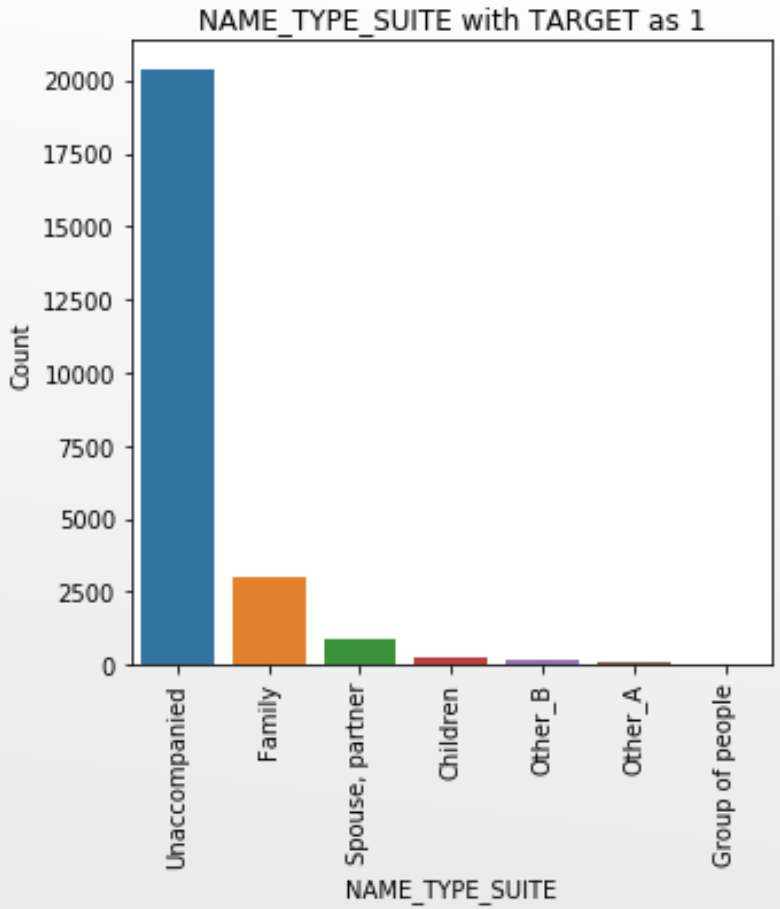


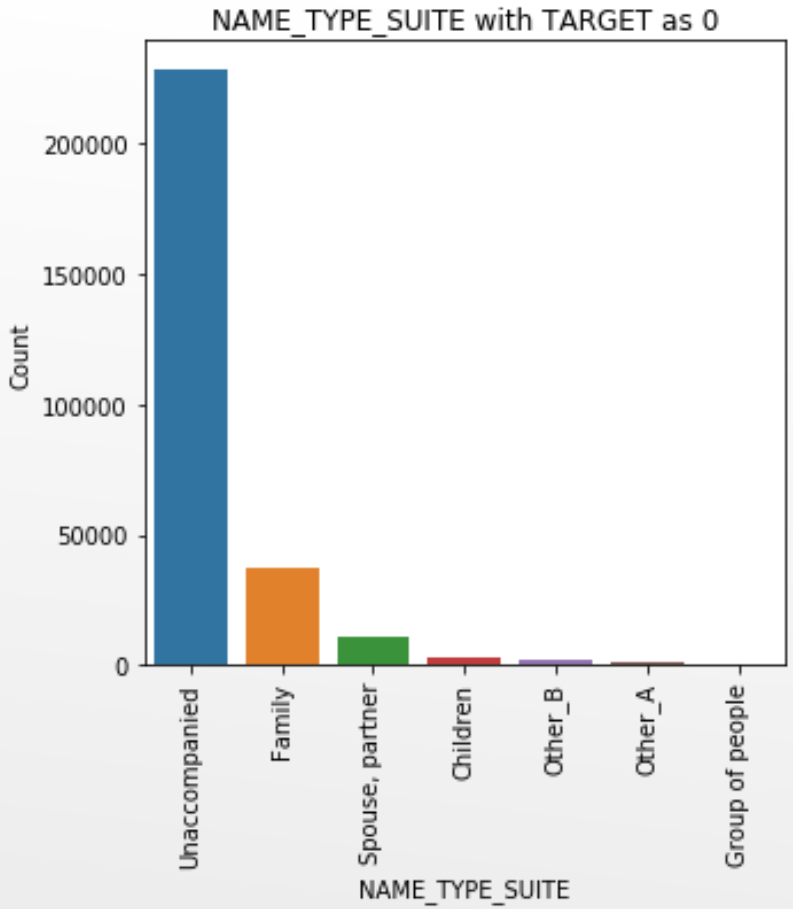
- Clients who do not own their own car opt for loans. We can deduce that they might go for loans for the same reason - getting their dream car.



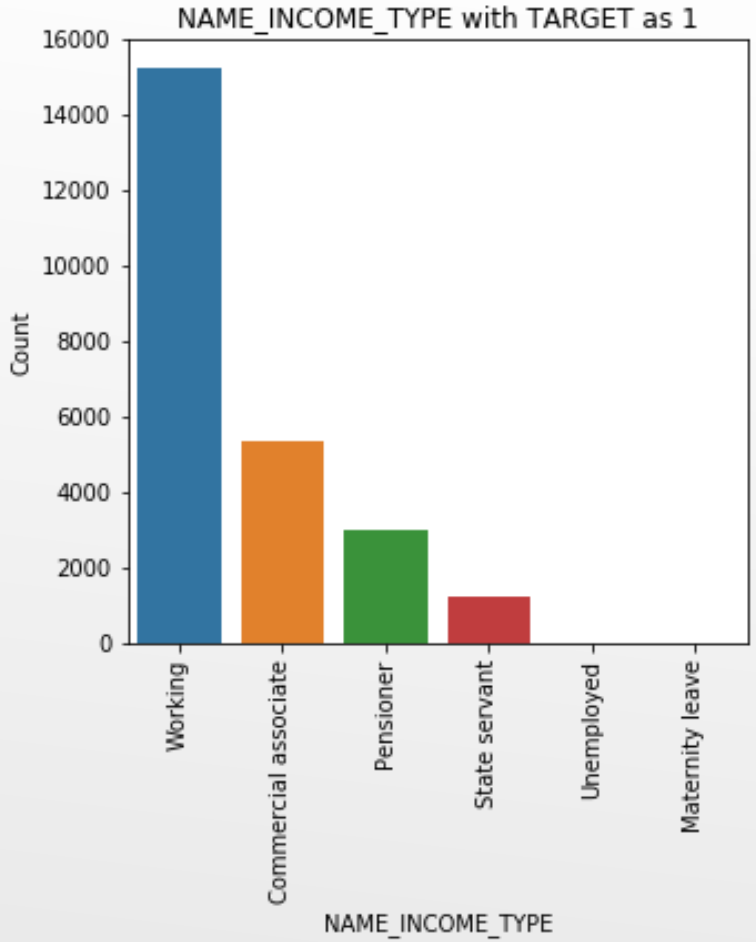


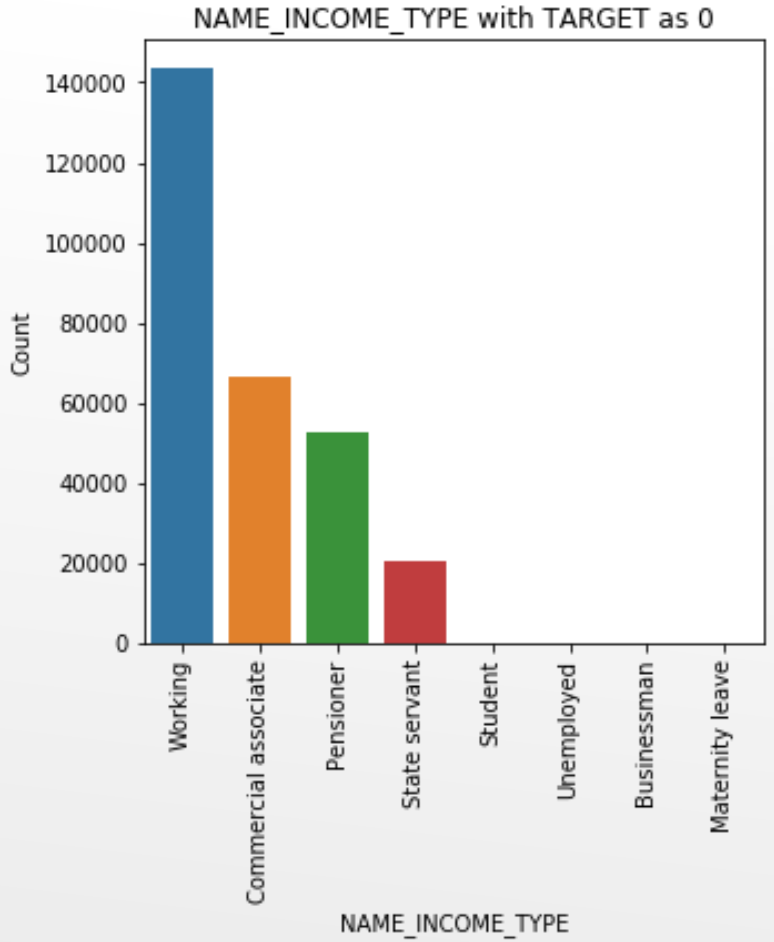
- Those who own their house/flat opt for loans more.



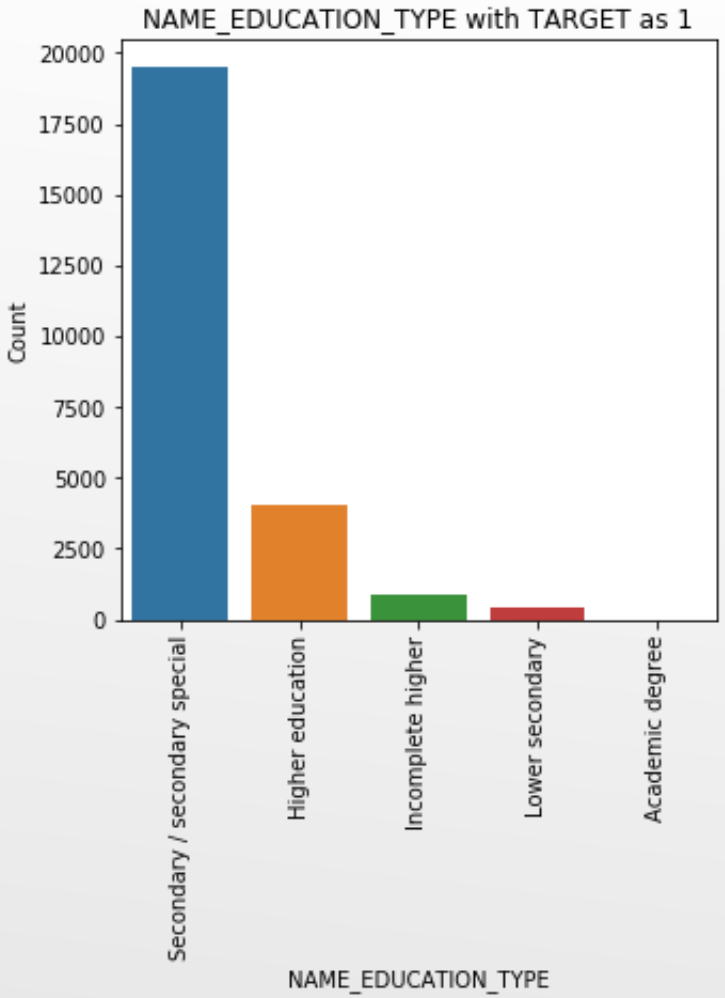


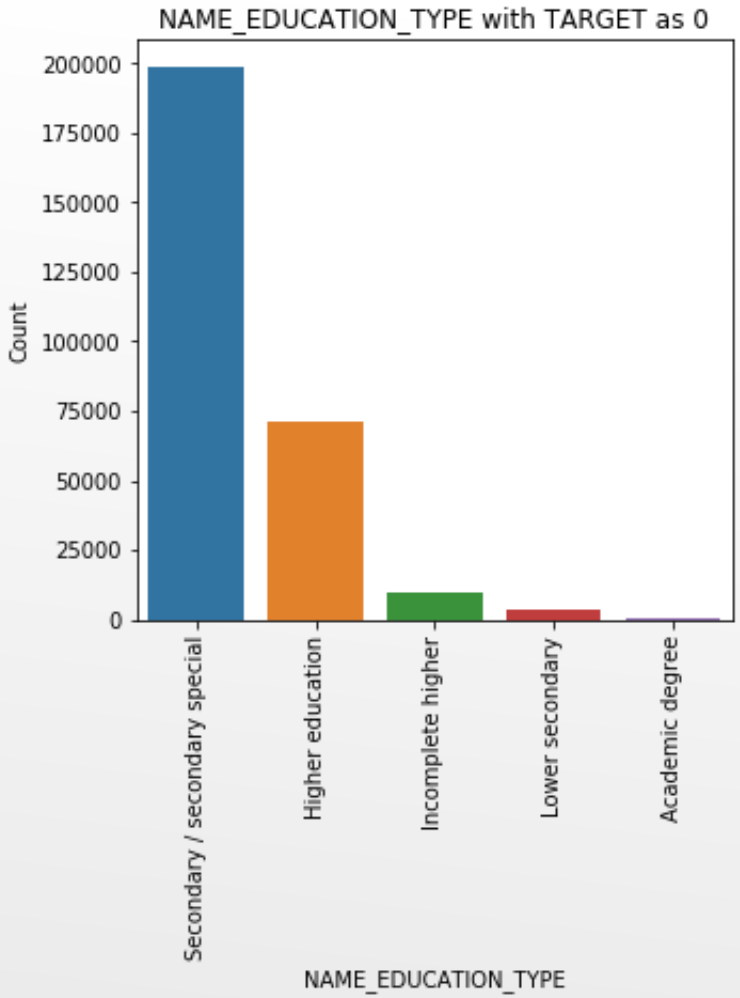
- Most of the clients who apply for loans are unaccompanied. And there is a drastic drop to the ones who are accompanied with their family, and so on with their spouse and children.



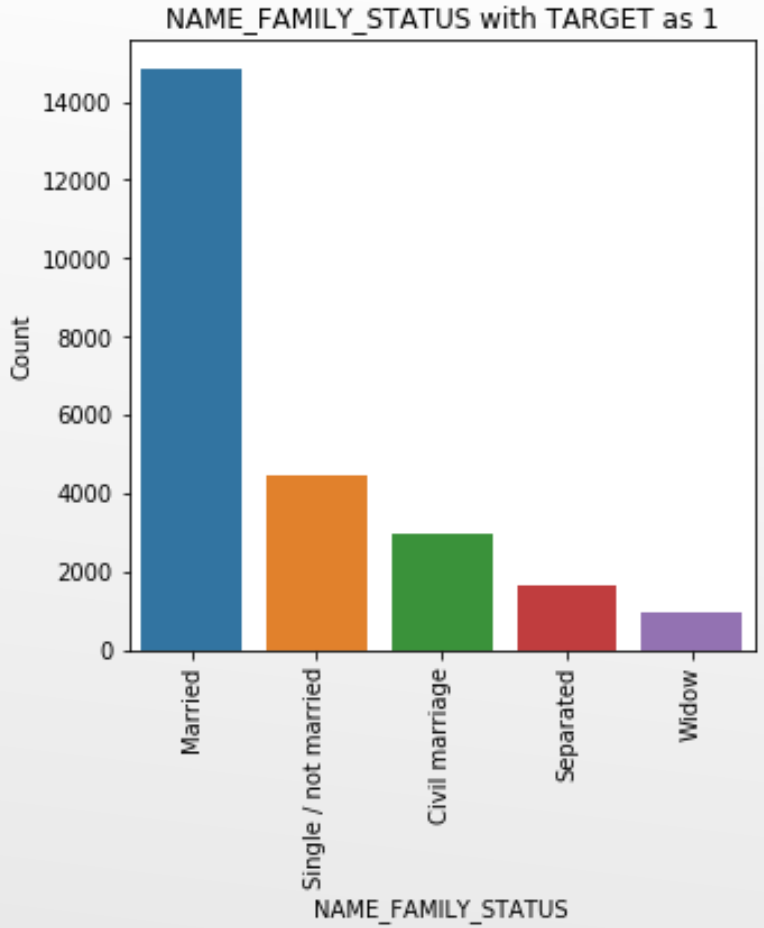


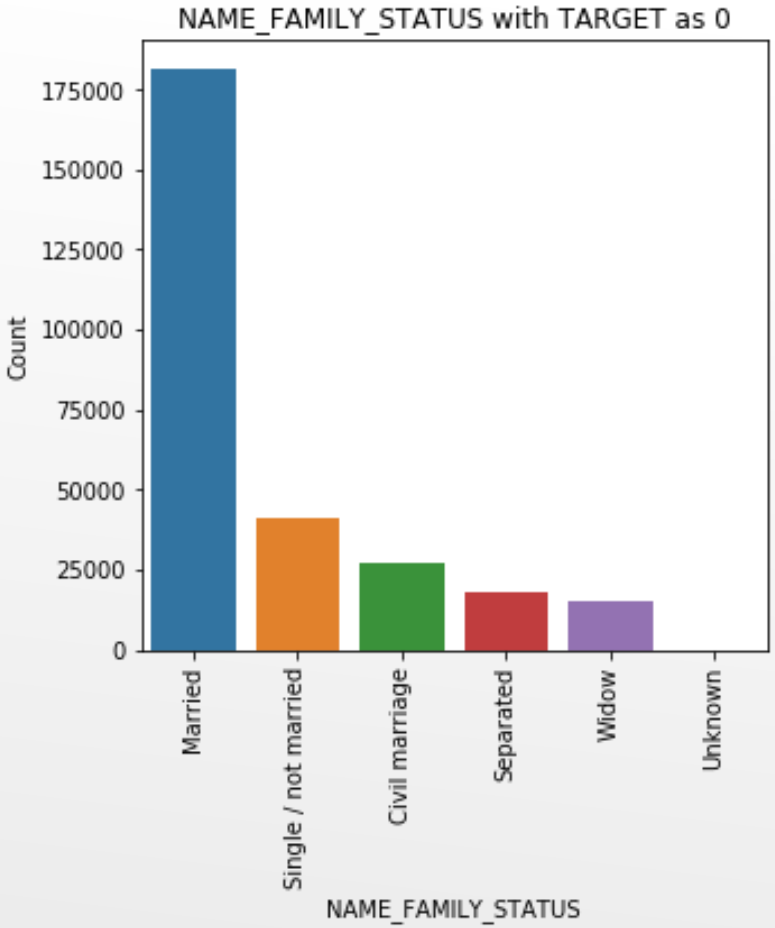
- Working people take more loans, approximately double to that of Commercial associates.
- State servants, unemployed and those on maternity leaves do take loans, but are very rare in numbers.





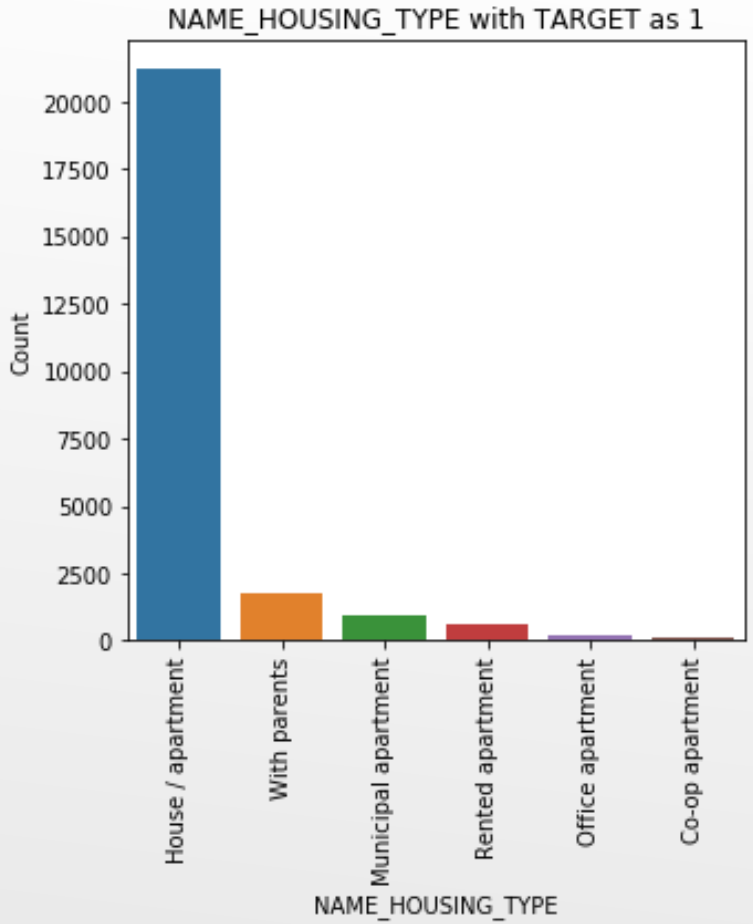
- Secondary educated clients go for more loans while those with academic degrees are very less in count to apply for the loans.

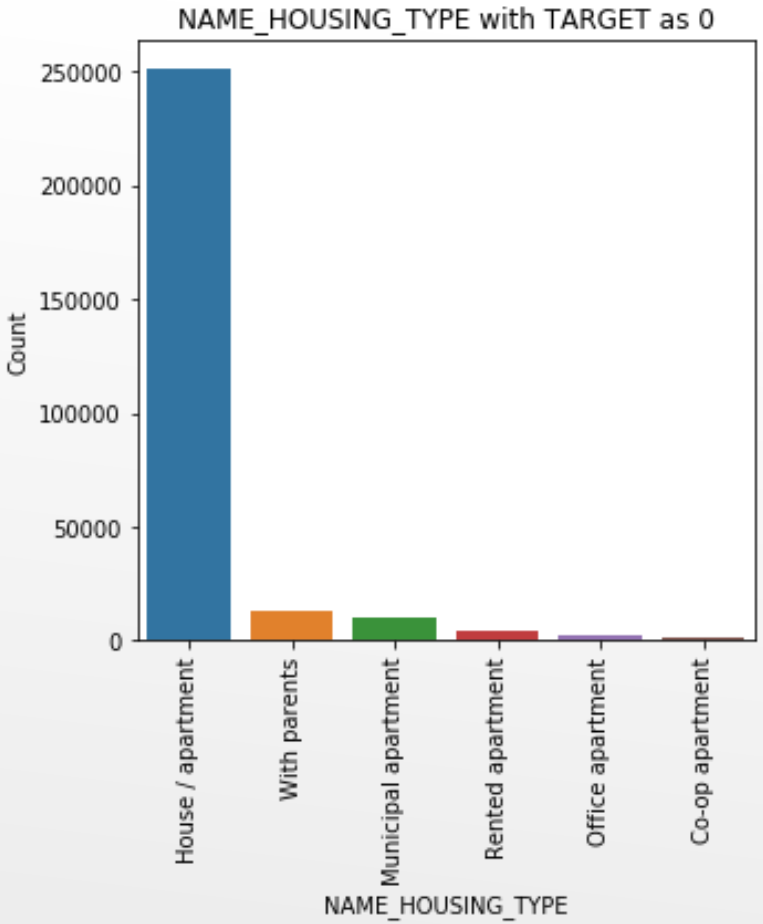




- Married clients apply for more loans followed by the singles and widows are the least to apply for the loans.

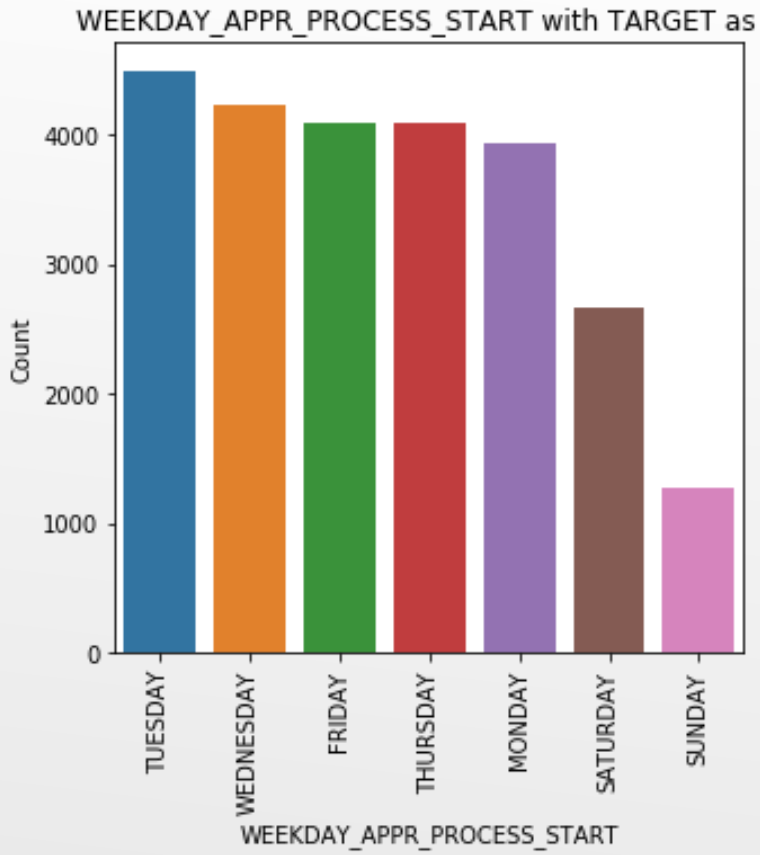
- This might be because marriage brings responsibilities onto the clients and might be at widowhood clients have accumulated the pensions and funds to survive. Widow clients might not have need for any expensive item.

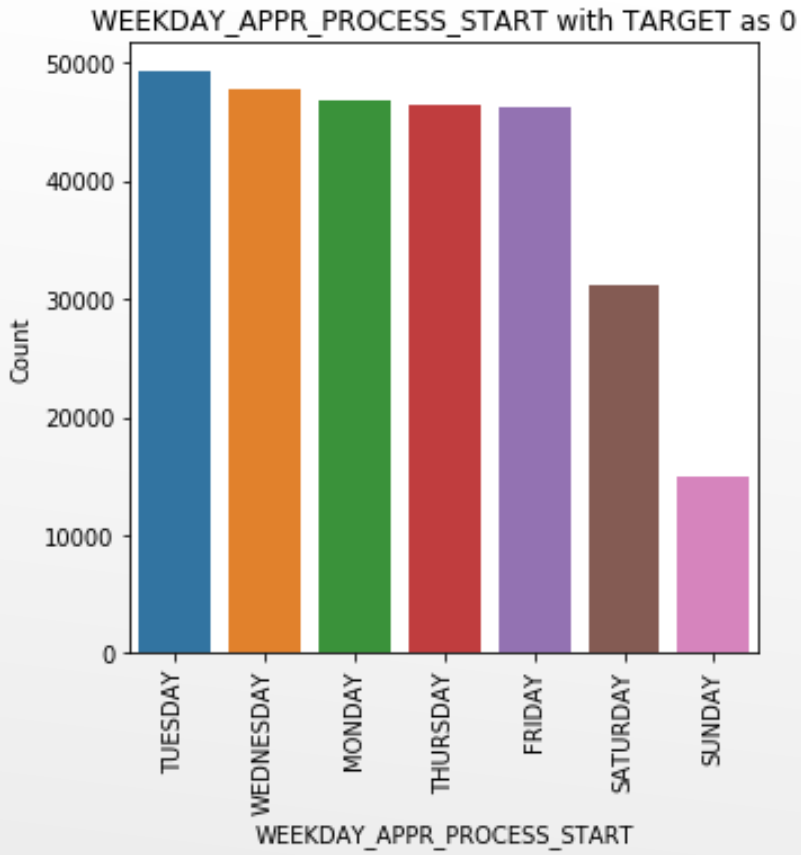




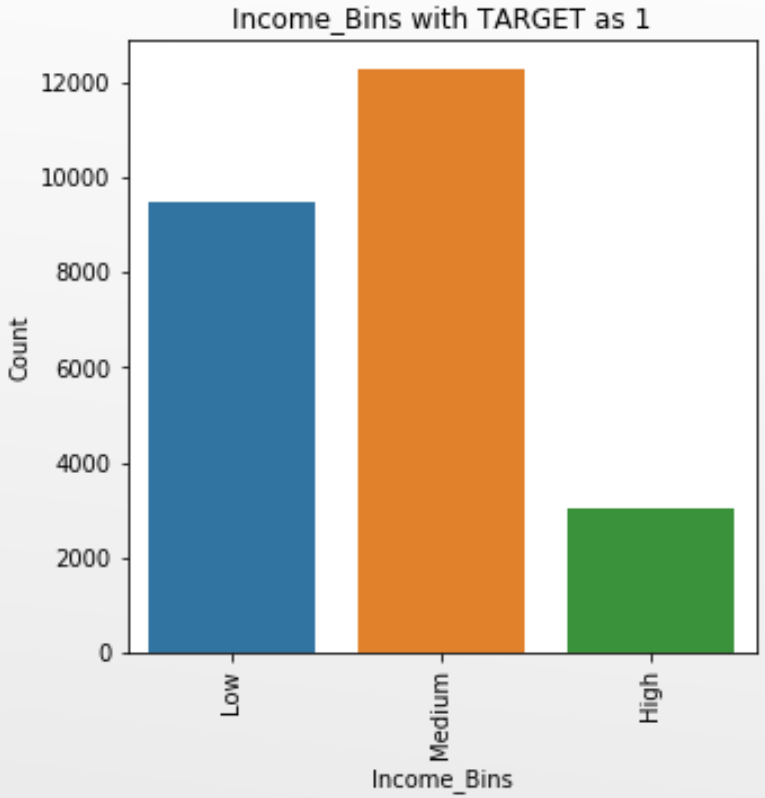
- It is weird that those who live in house/apartments apply more for the loans while those on rented ones do not.

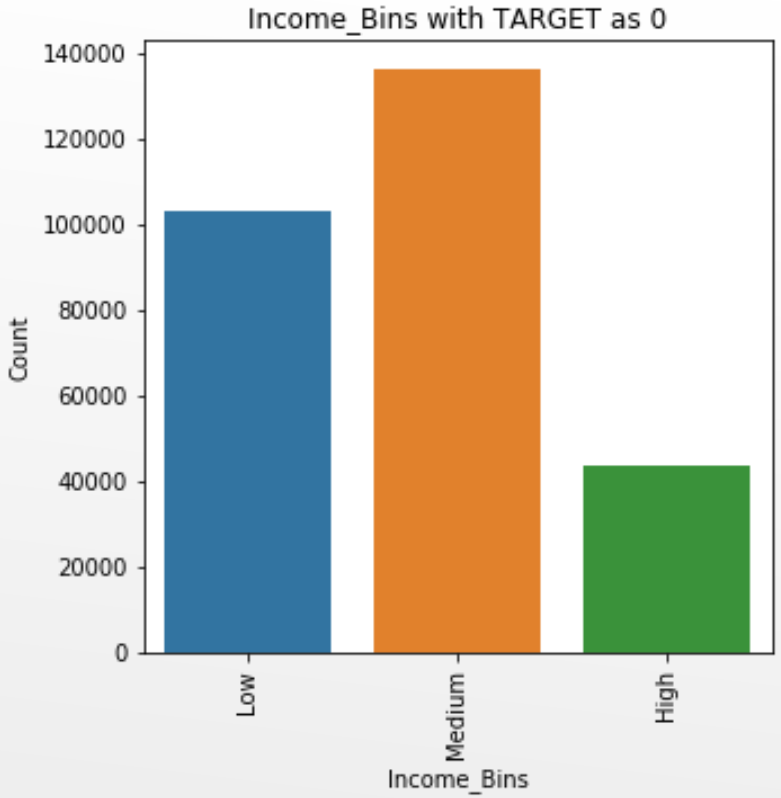
- Might be those who have house/apartment are nuclear families and since they do not stay with their parents have more desires for expensive items. To be precise more than **10 times** in comparison to staying with the parents.



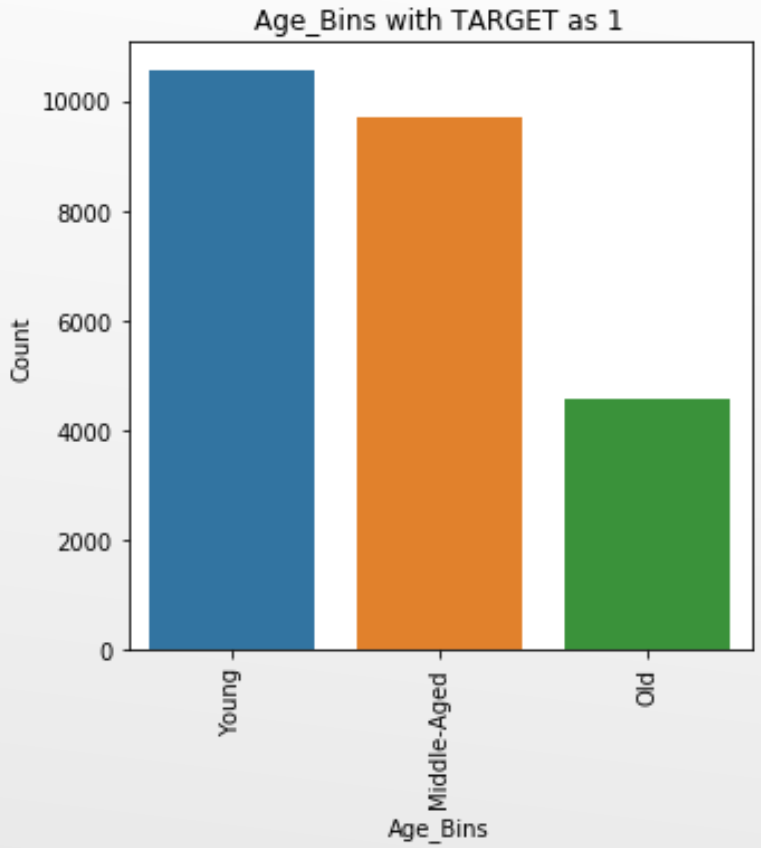


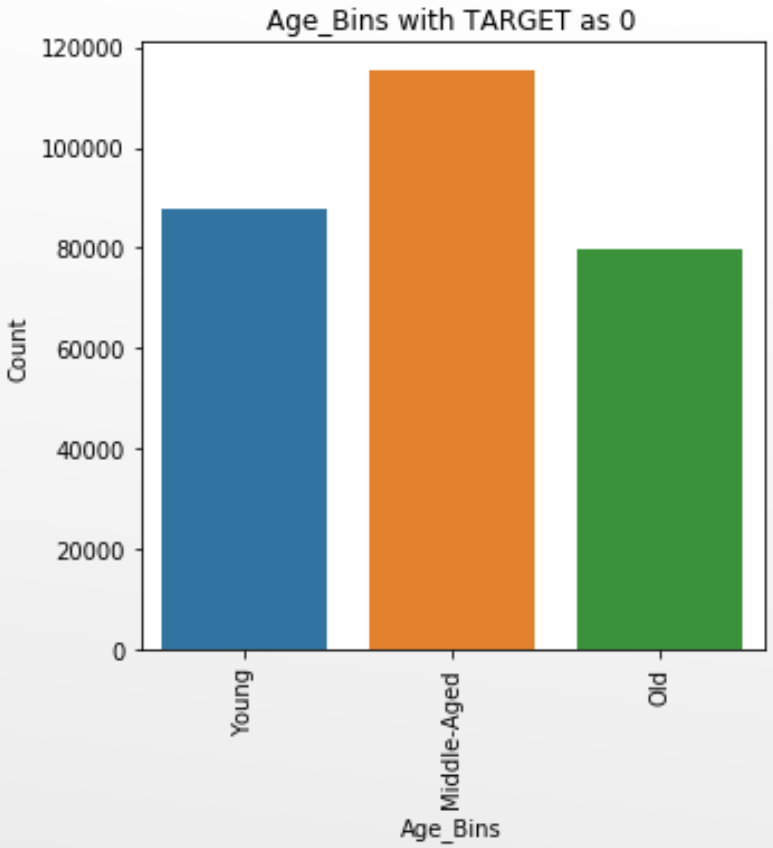
- It is interesting to note that clients with **TARGET 1** favor Fridays than the Mondays, which is just *opposite* in the case of the clients with **TARGET 0**.



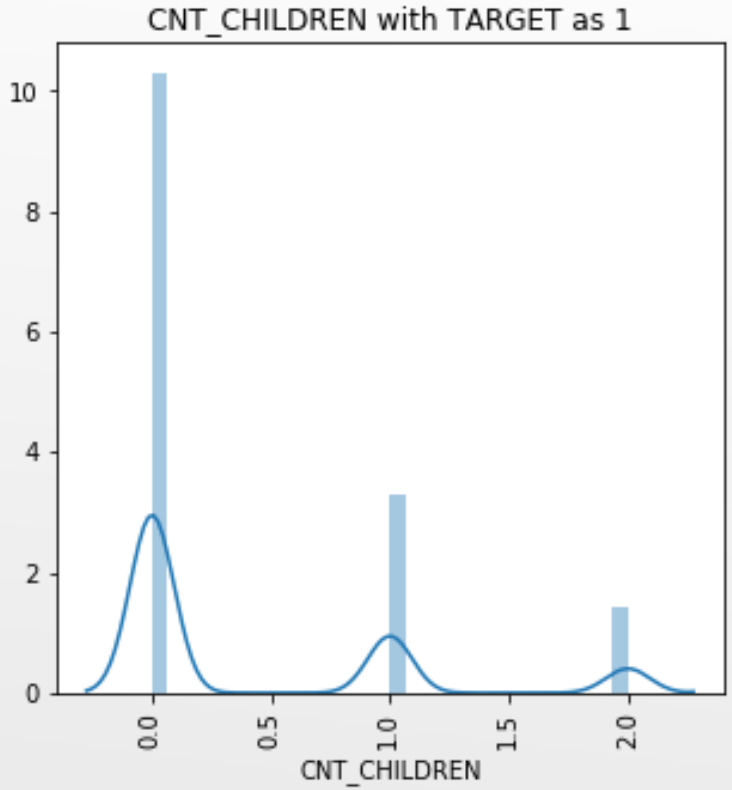


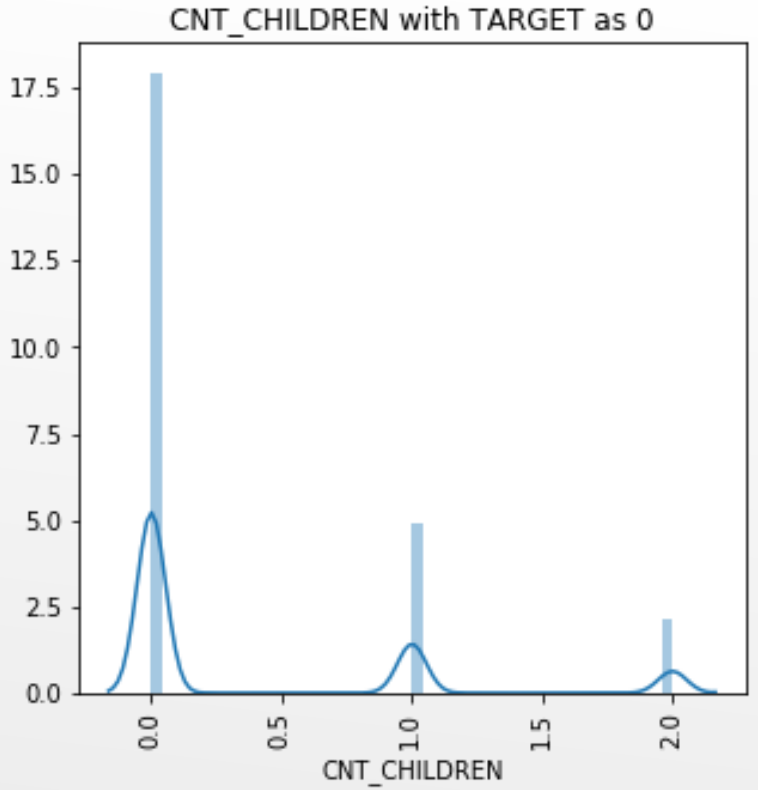
- Clients with Medium incomes apply for more loans than those with low incomes. The very obvious thing is that those with high incomes are less in numbers.



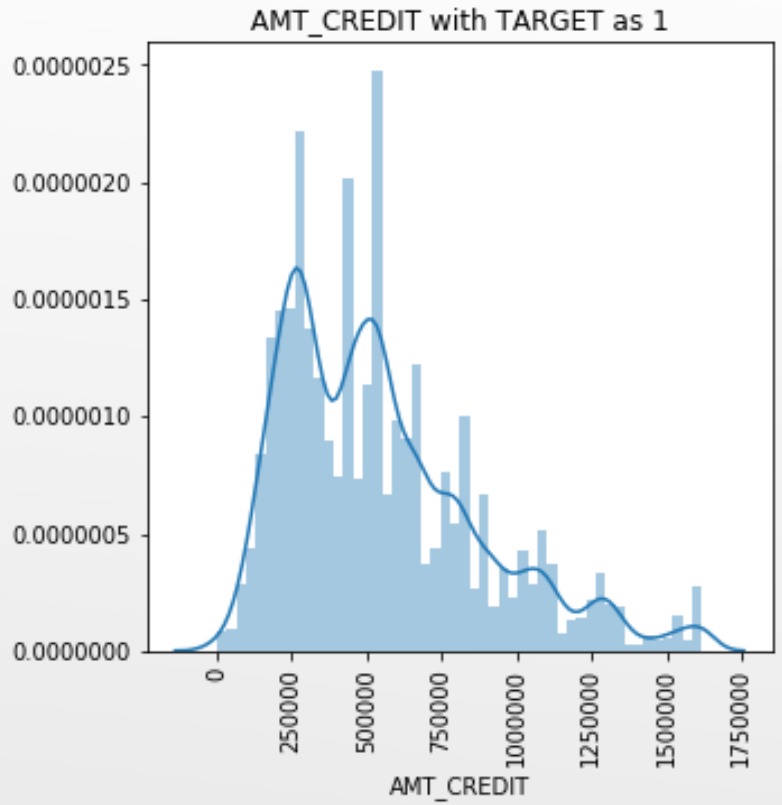


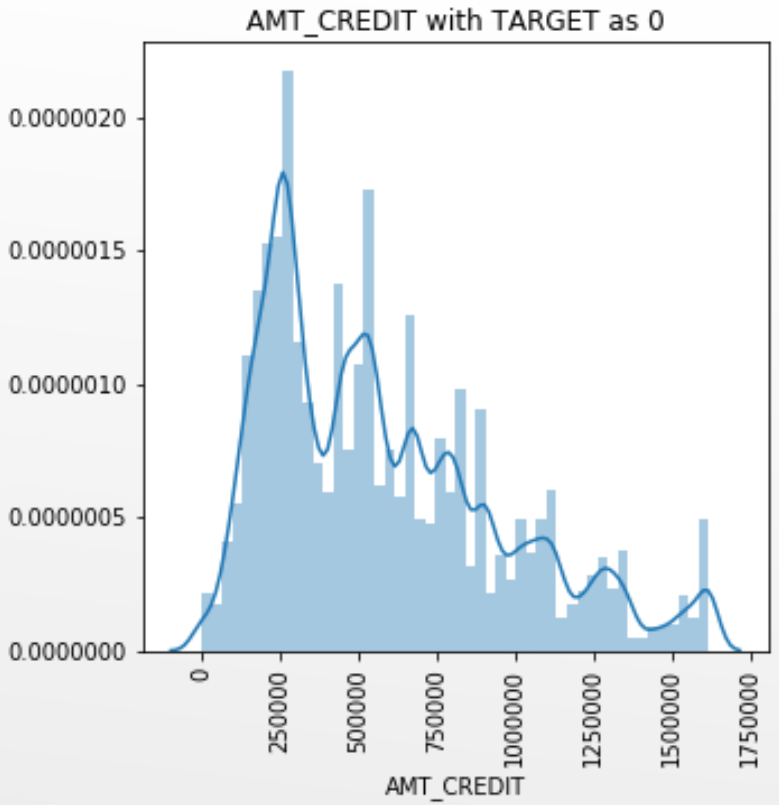
- For clients with **TARGET 1** young people are the most who apply for the loans, while for clients with **TARGET 0** middle-aged are the ones more to apply for the loans. That is a very crucial factor to decide the defaulter.



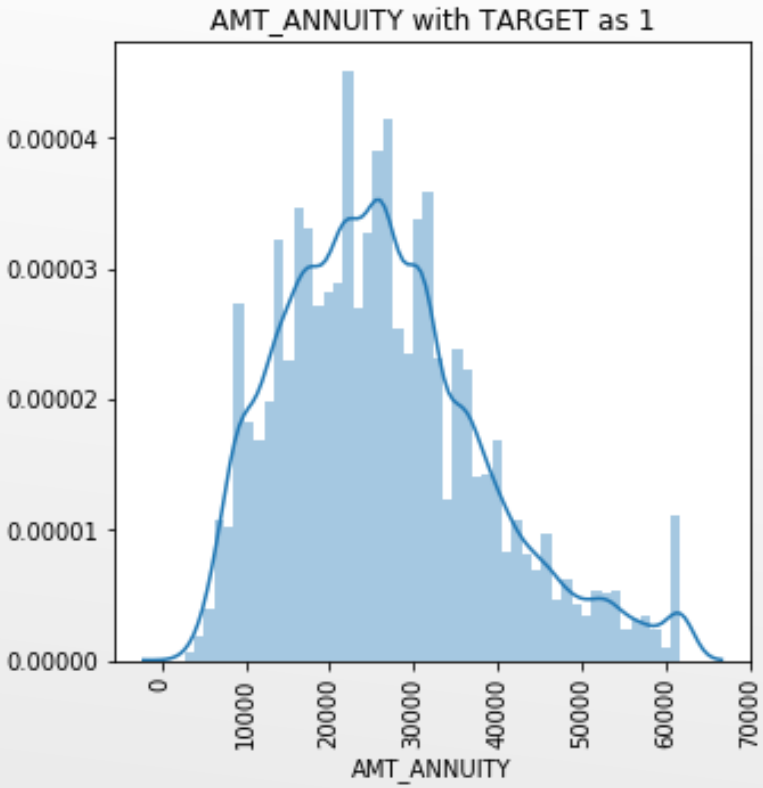


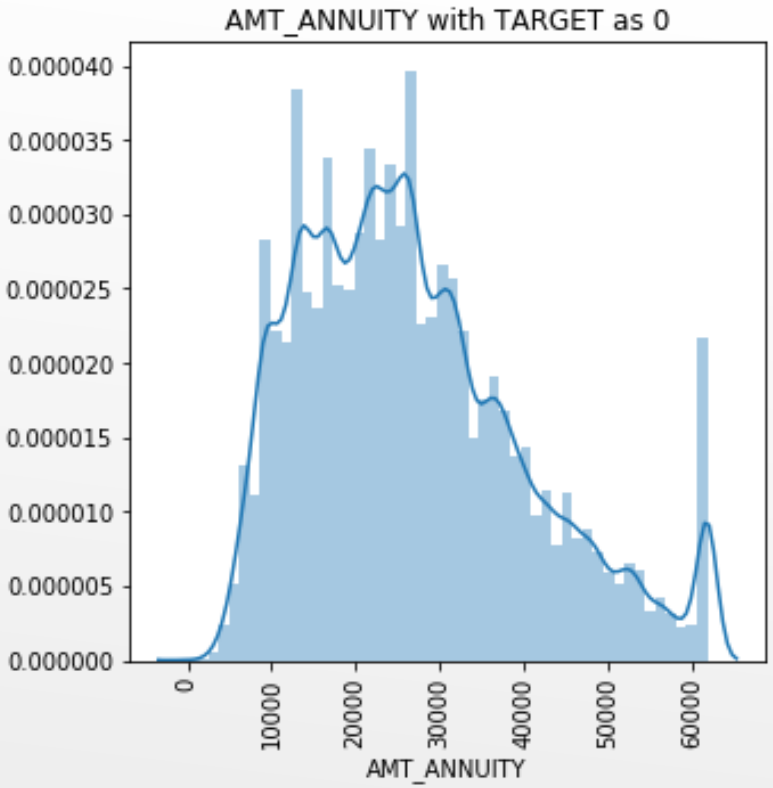
- It is 5 times more probable for a client be having **no child** in comparison to those with **2 children**. This signifies that more clients who opt for loans bear **no children**.



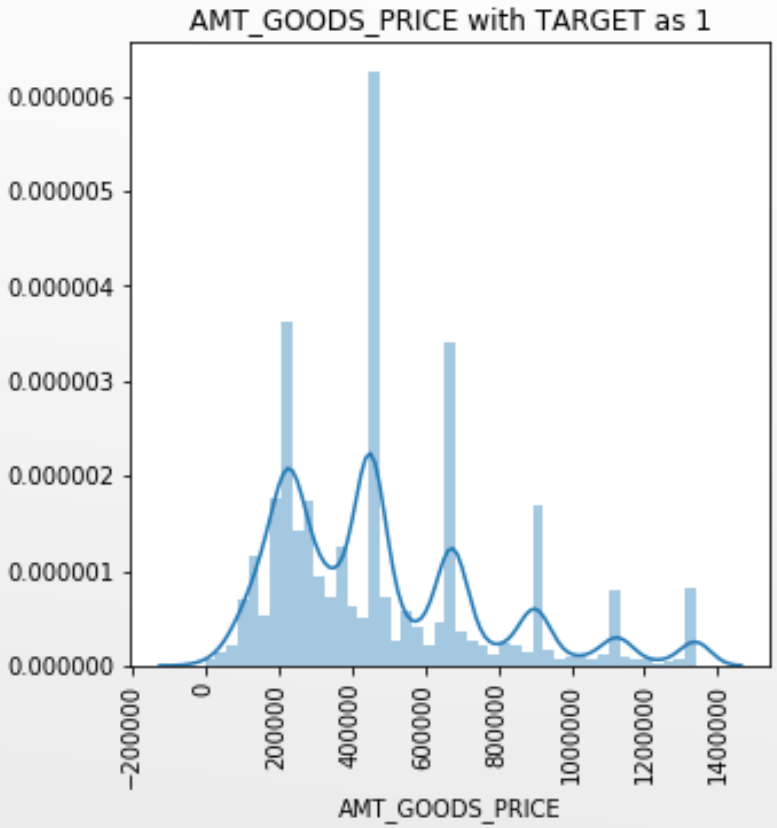


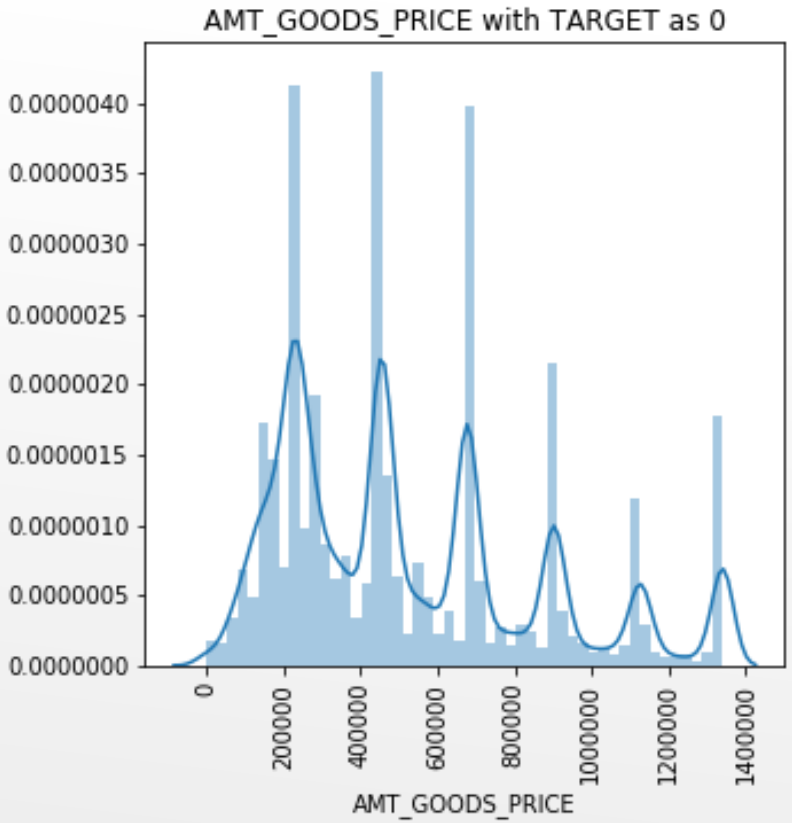
- We can see clients (regardless of the TRAGET category) opt for **loans** amount **2,50,000** and **5,00,000**. Also an interesting thing is that there is a sudden spike at **16,50,000**.



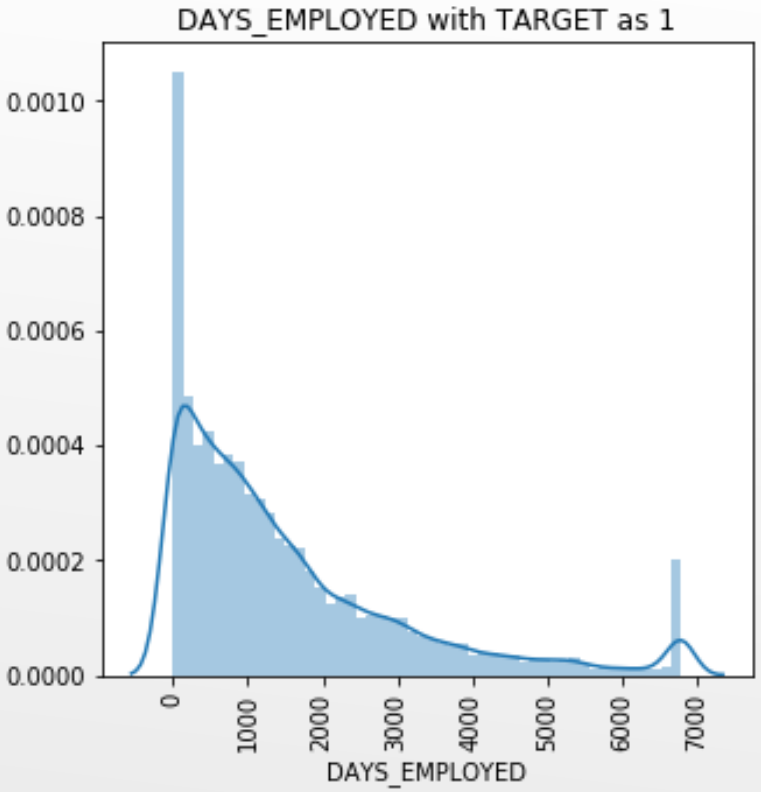


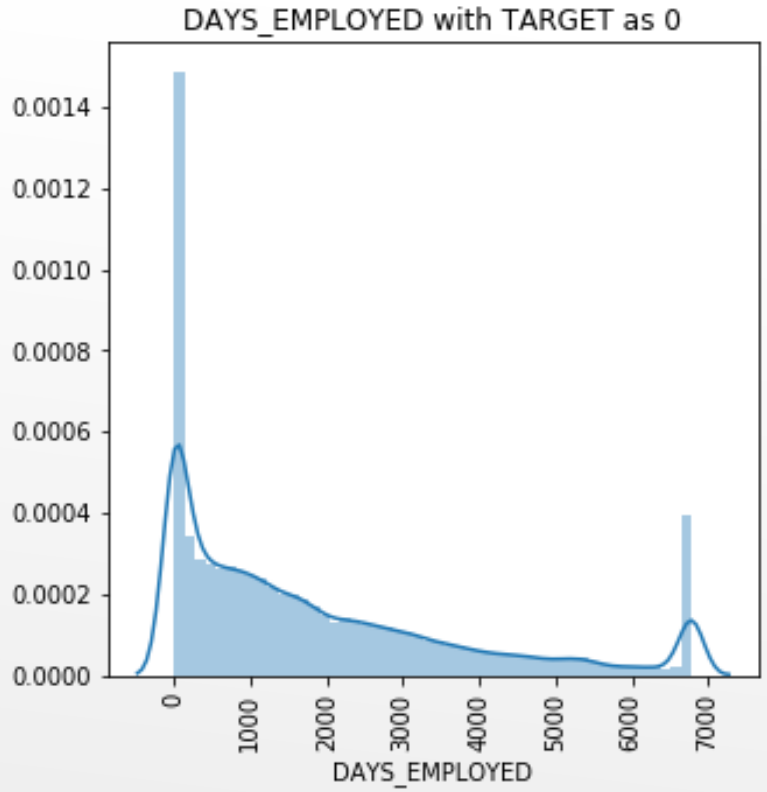
- As seen in loan credit amount, there is a similar spike in **annuity** amount at **60,000.**



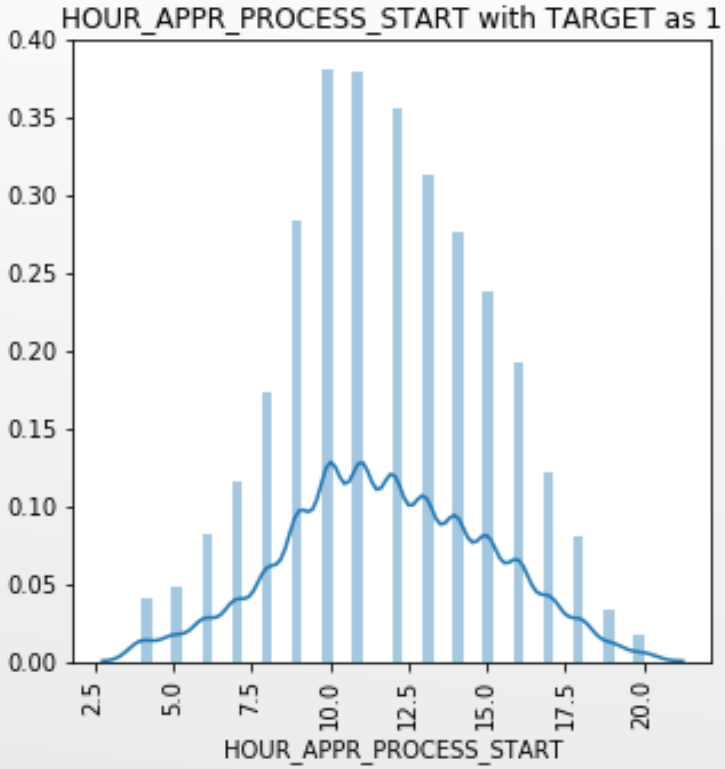


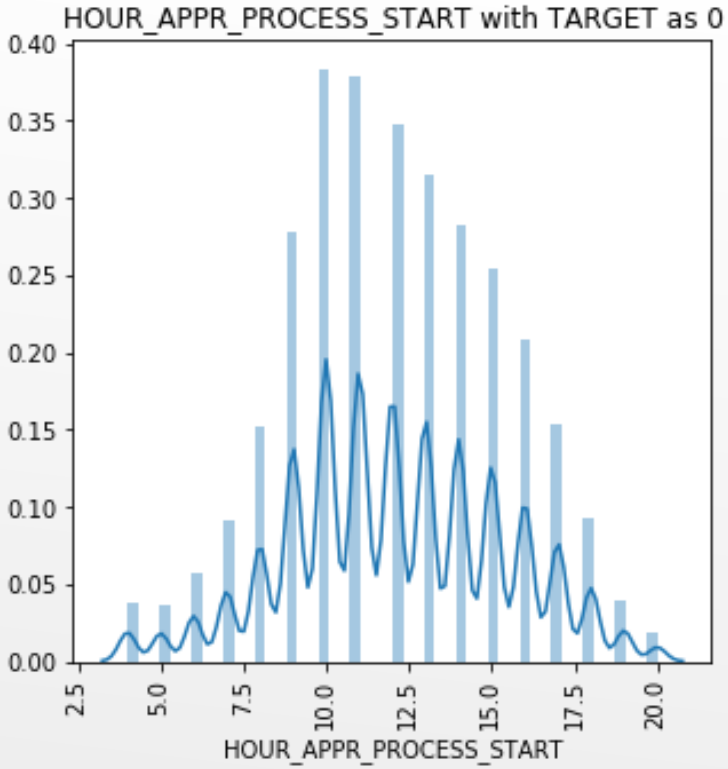
- Clients who belong to **TARGET 1** have more loans for goods amounting of **4,00,000** in comparison to **2,00,000 and 6,00,000** while those with **TARGET 0** have **equal** probability densities for the goods they take loans for.



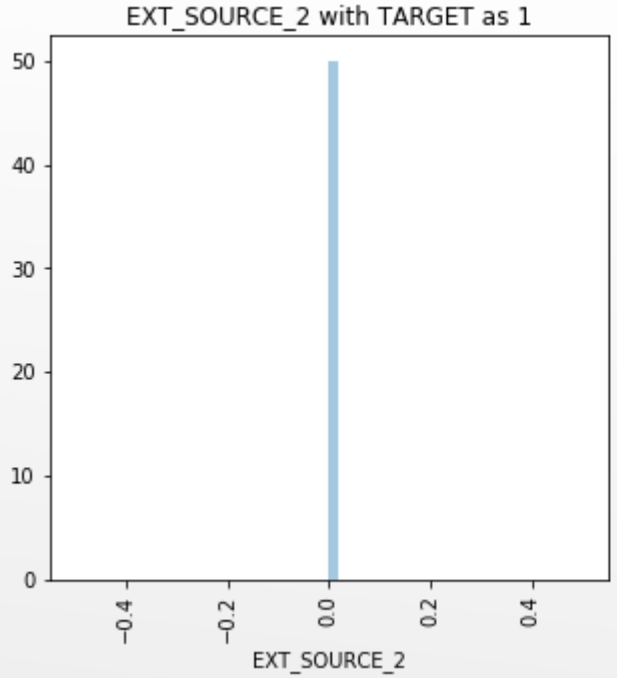


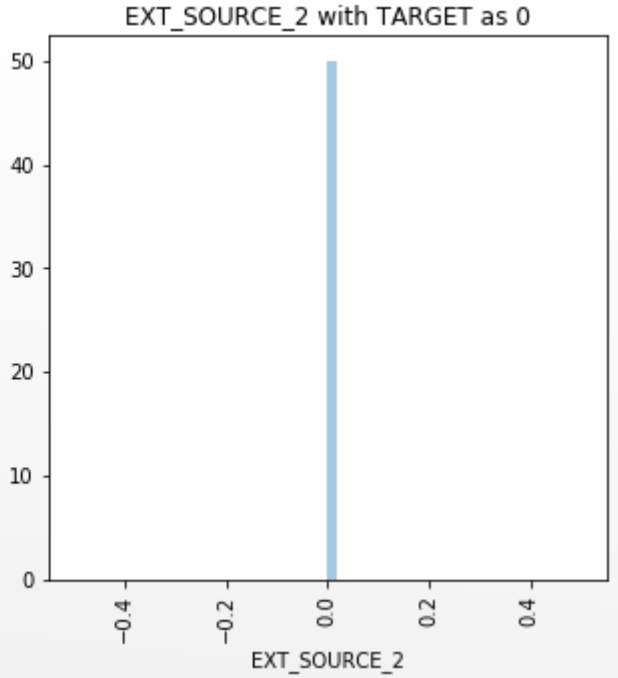
- One interesting thing to note is that clients about **19-20 years of experience** have shown sudden spikes for opting for the loans.

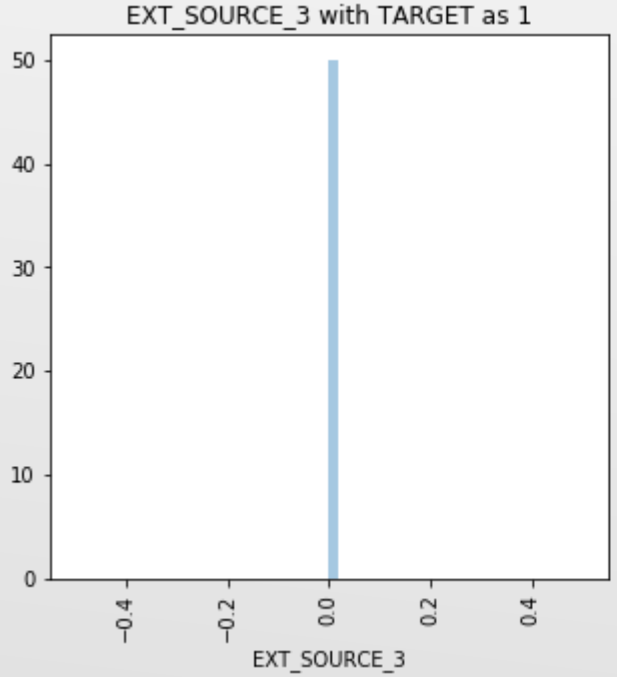


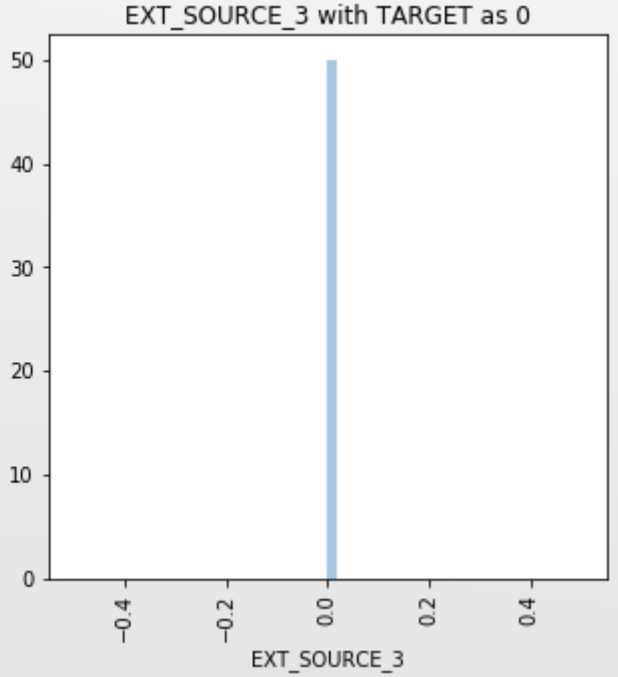


- Most applications have been filled for early office hours like **10 AM and 11 AM.**

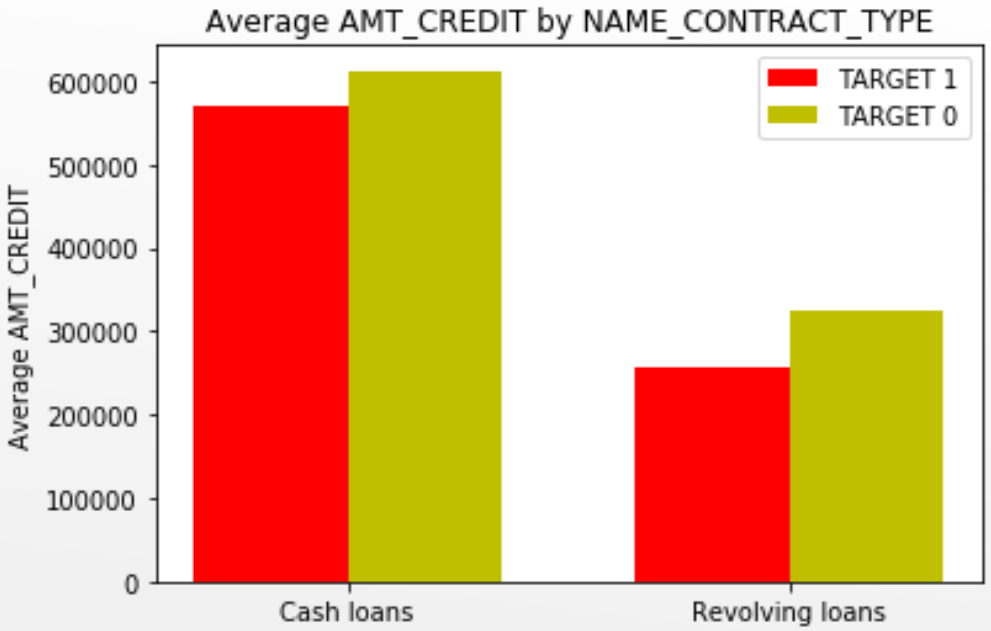




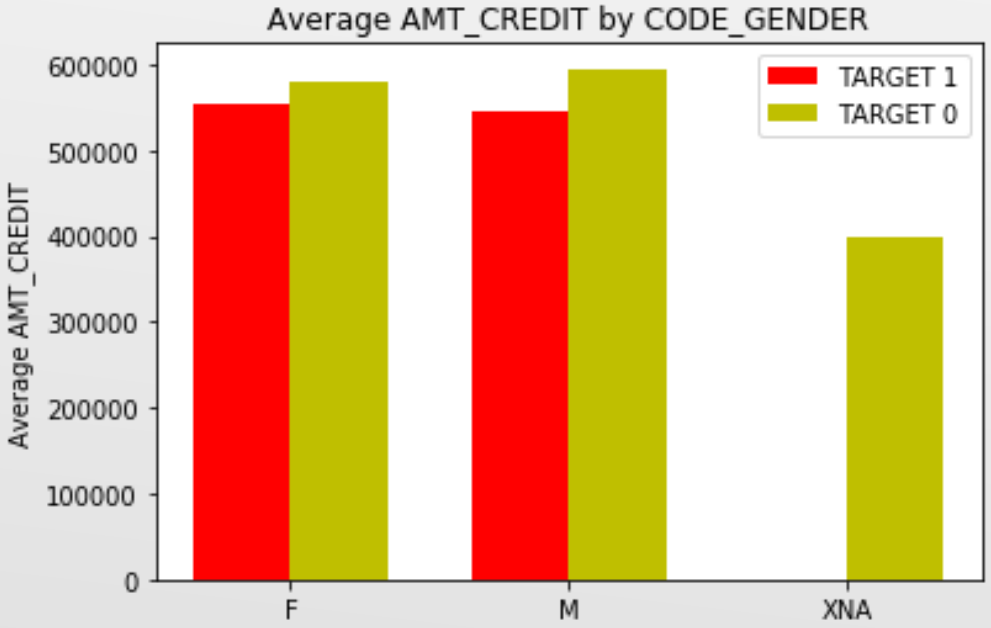




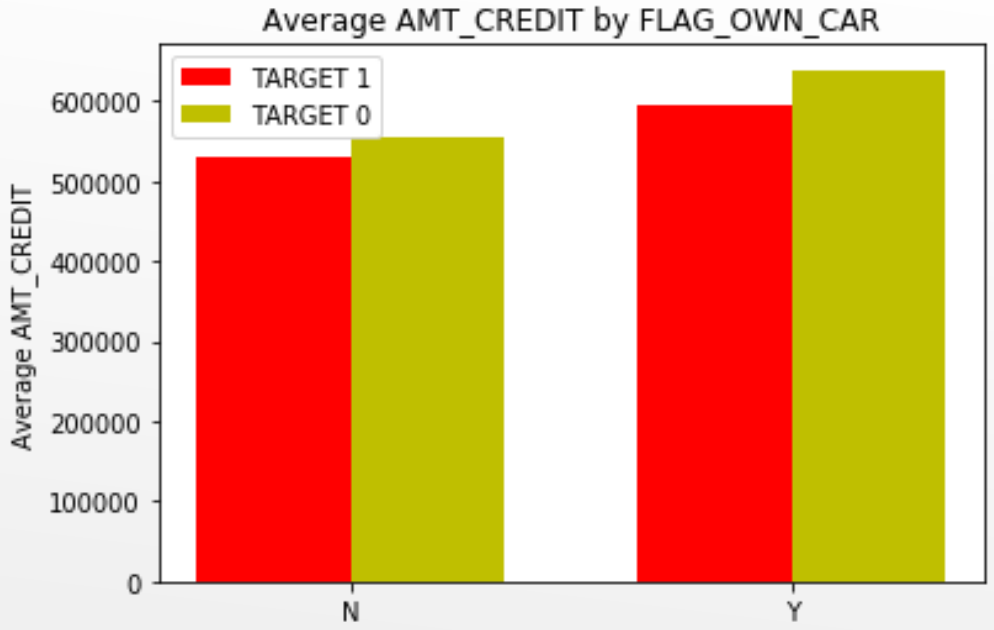
- The exterior sources have **negligible difference** for the scores.



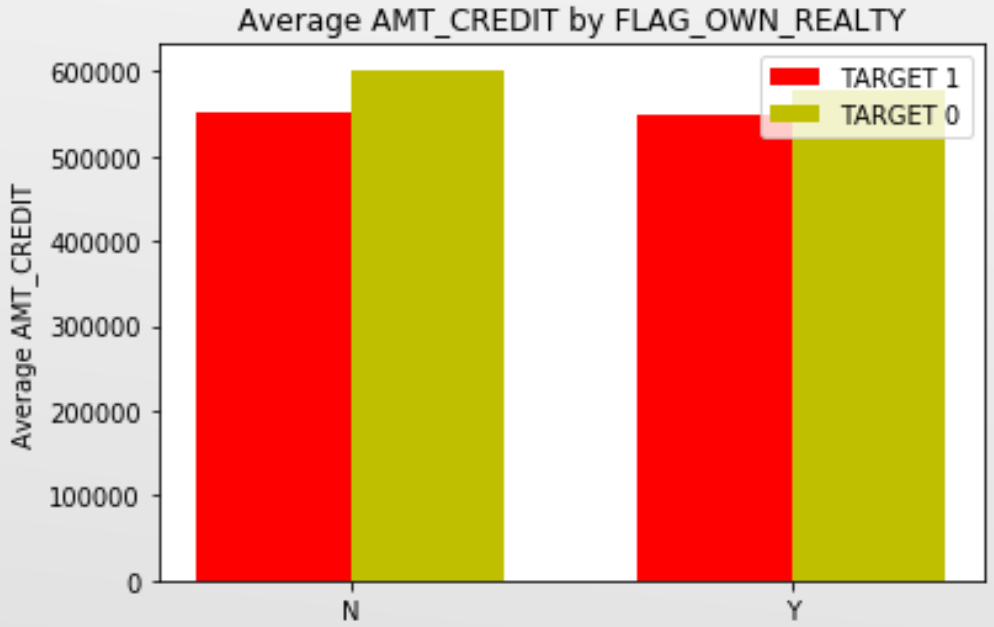
- The average credit amount for any kind of loan (**cash loans or revolving loans**) have been a little high for **TARGET 0** than **TARGET 1**.



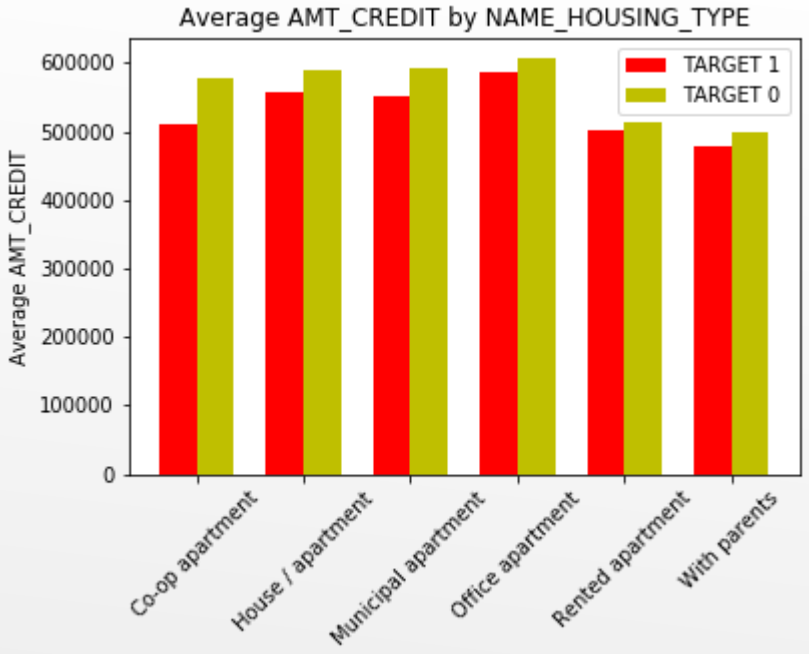
- Seeing the average credit amount for **Females** and **Males** have been found as mostly **similar** for any of the TARGET categories.
- Those who hide their gender (**XNA**) mostly do not face any difficulties in paying back the loans but their count is very less.



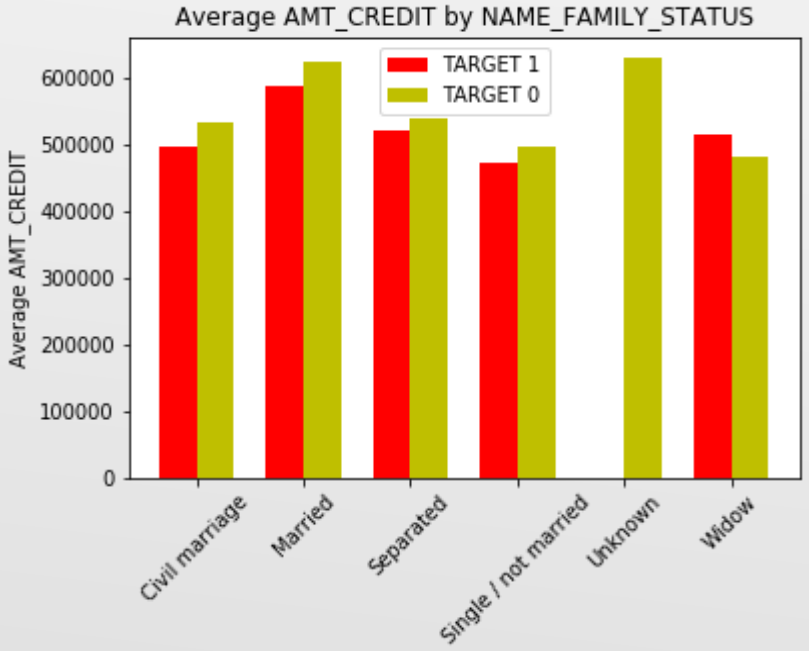
- The average credit loan amount is slightly less for **TARGET 1** than TARGET 0 when considering whether the clients **own the car or not**.



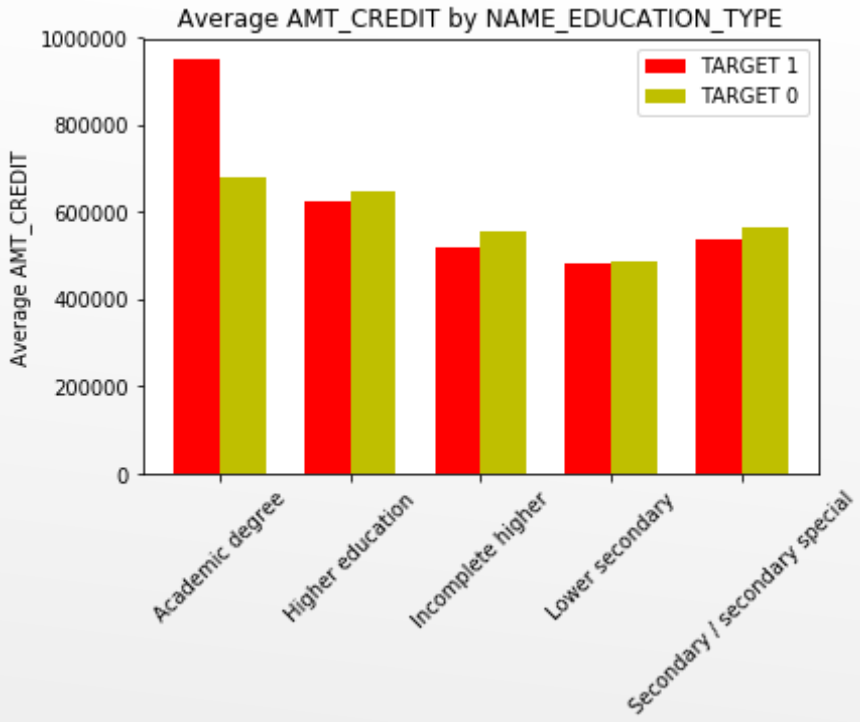
- The average credit loan amount is slightly less for **TARGET 1** than TARGET 0 when considering whether the clients **own the house/apartment or not.**



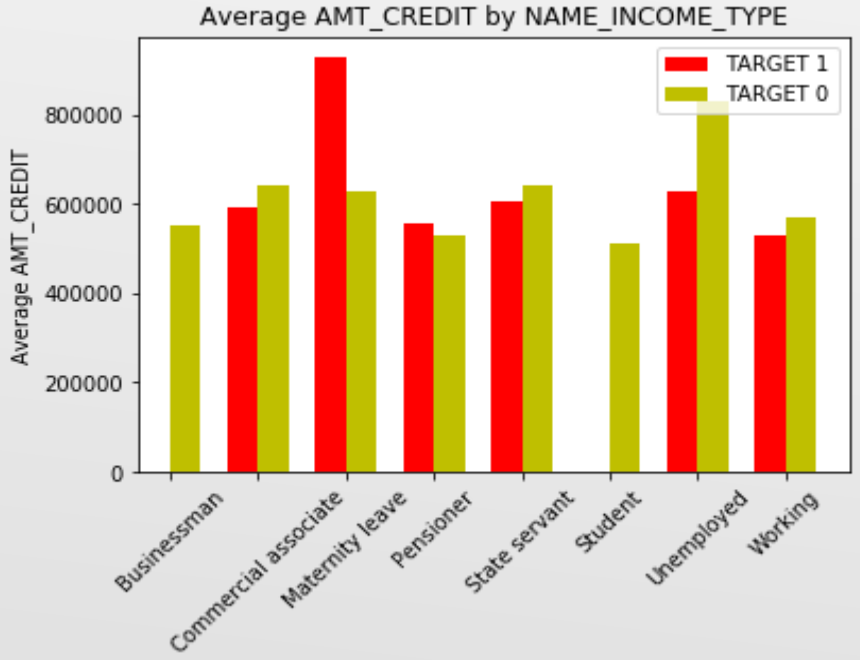
- Let it be client staying in any kind of home (**co-op/own their house/municipal**), their average credit amount is higher than those staying with **parents or in rented apartments**.



- We notice for clients who are **married/single/separated**, the average credit amount for **TARGET 1** are less than those for **TARGET 0**. But the striking thing is that this is opposite for **Widows**.

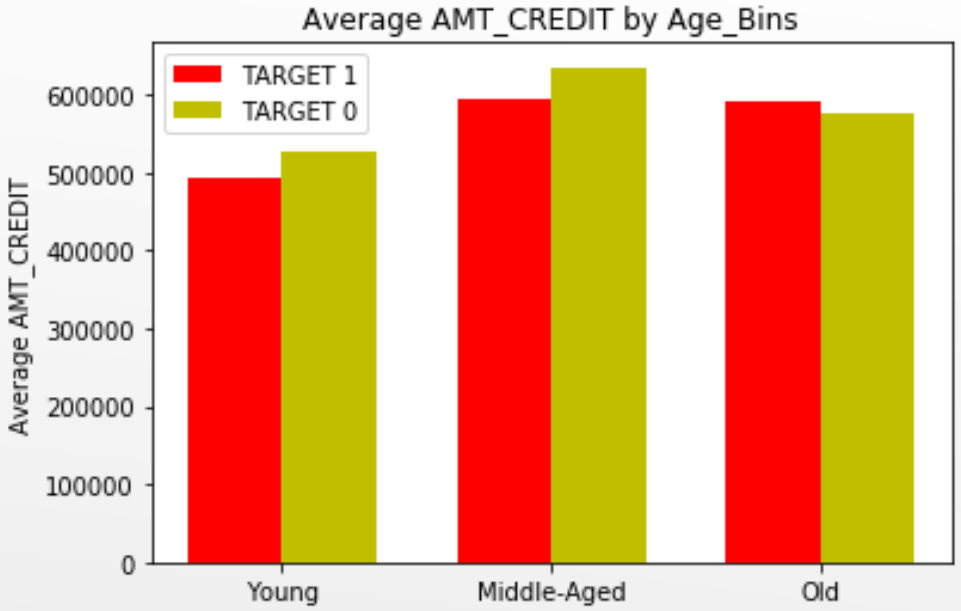


- That's very striking to notice that those with **academic degrees** take more credit loan amount and fall in **TARGET 1**. This has been very opposite to those with **lower secondary**.

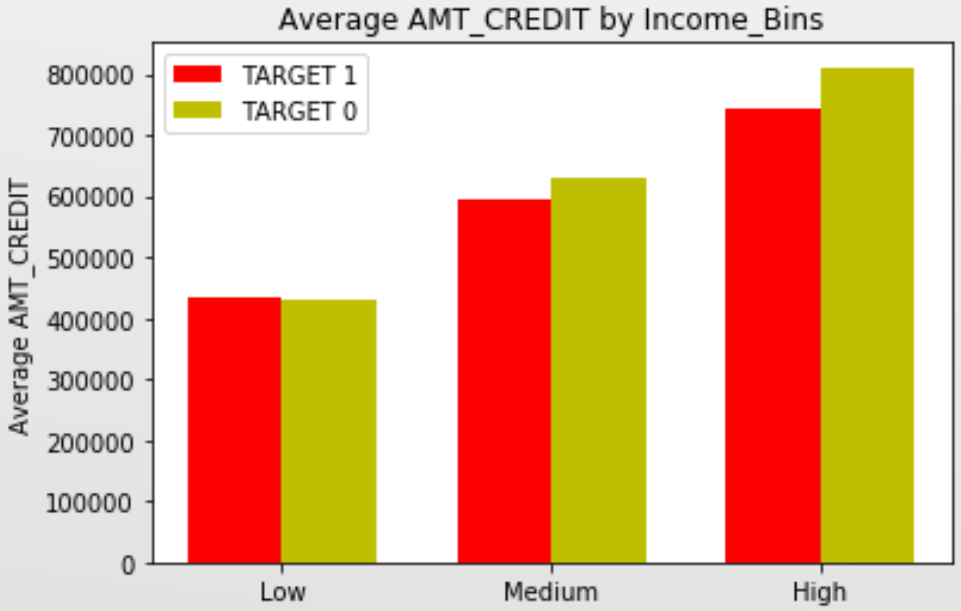


- **Businessman** never face any difficulties in paying back the loans. That sounds obvious. Even **students** do not face in paying back the loans.

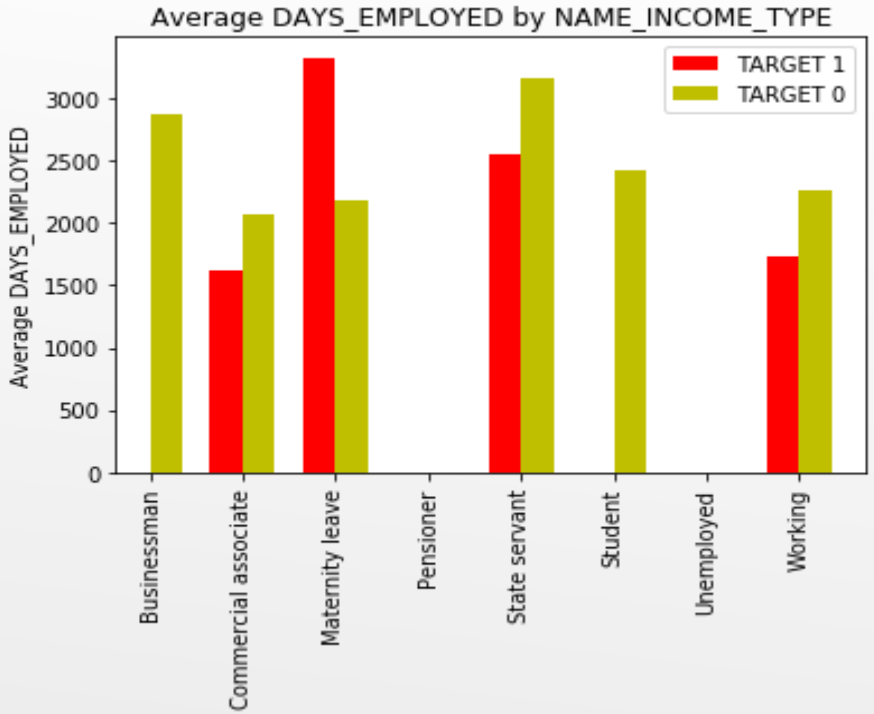
- But the credit amount taken by people on **Maternity** leaves is susceptible to get defaulted while that is not the case with **unemployed**.



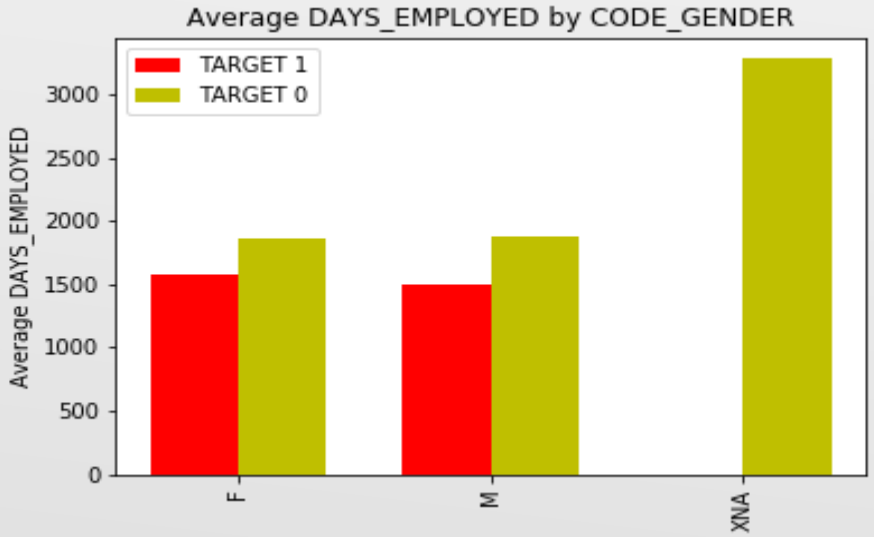
- Clients in **Old** age are more likely to be defaulter with high average credit amount than those in **TARGET 0** category.



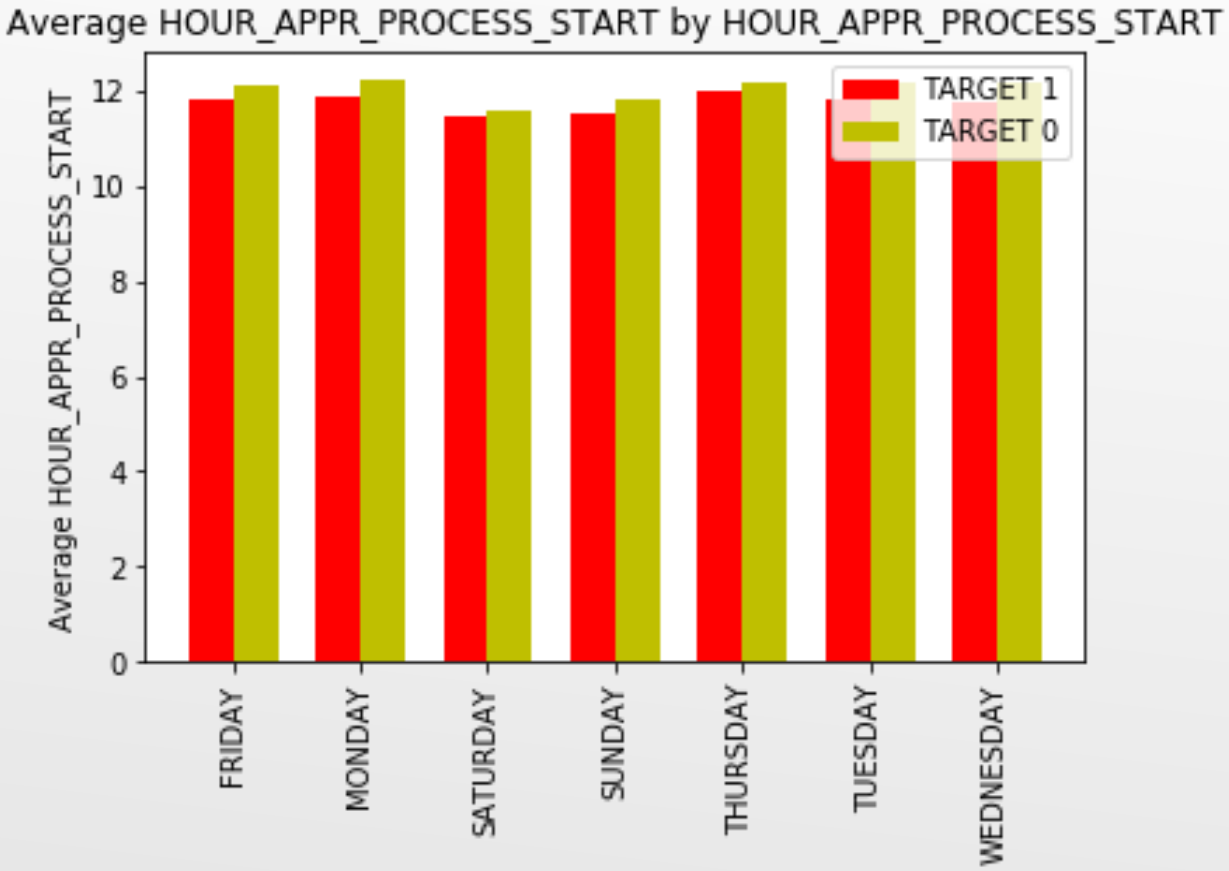
- Average credit loan amount increases as the **income** increases of the clients.
- But for clients with **low incomes**, the defaulters take more loan amount as compared to **TARGET 0**.



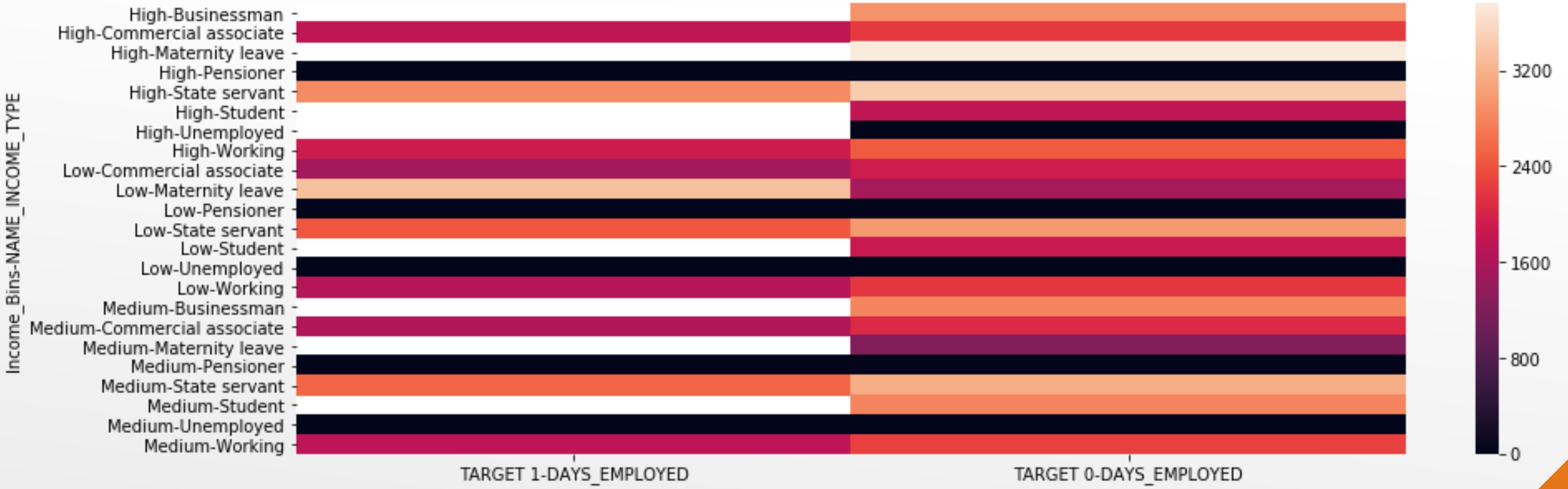
- Clients on **maternity** leaves have higher average employment days with **TARGET 1** than compared to **TARGET 0**.



- Average employment days for any gender (**Female or Male**) for **TARGET 1** category is less than **TARGET 0** category.

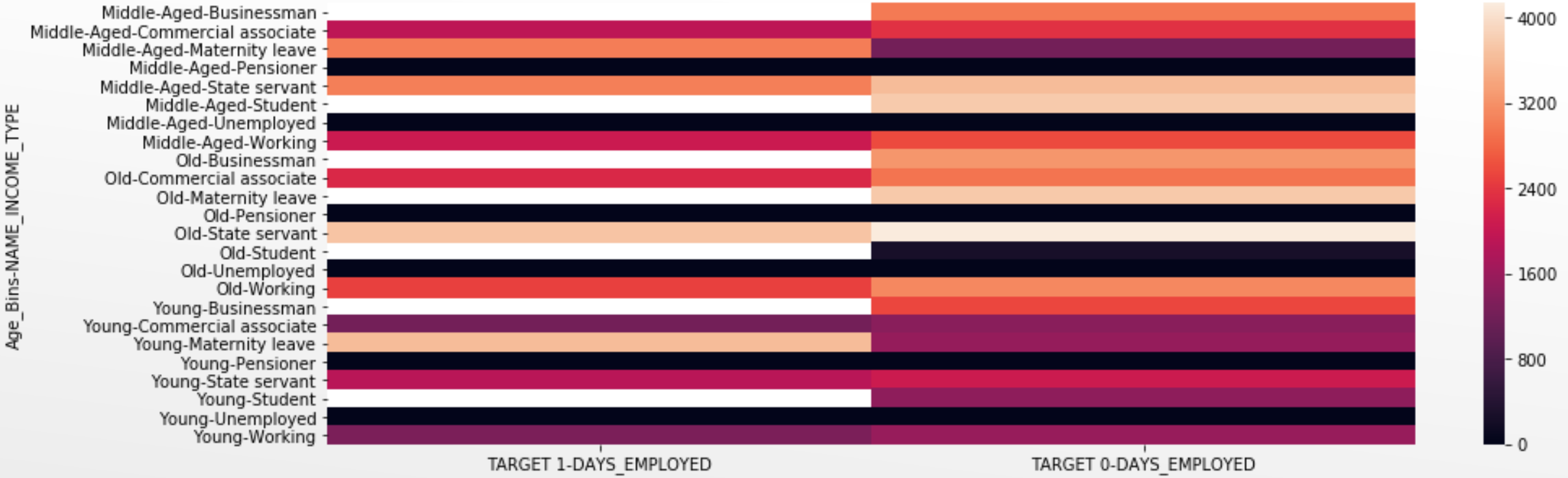


- Loan application starting process is generally seen at 12 PM in the afternoon, let it be any day.
- To be precise, **TARGET 1** come slightly earlier on any day for loan applying than **TARGET 0**.



- We see incomes for most of the various income types is less (as more purple than orange) for **TARGET 1** than that in **TARGET 0**, considering the average number of employment days.

- That directly tells that Income Brackets matter for the client being a defaulter or not.

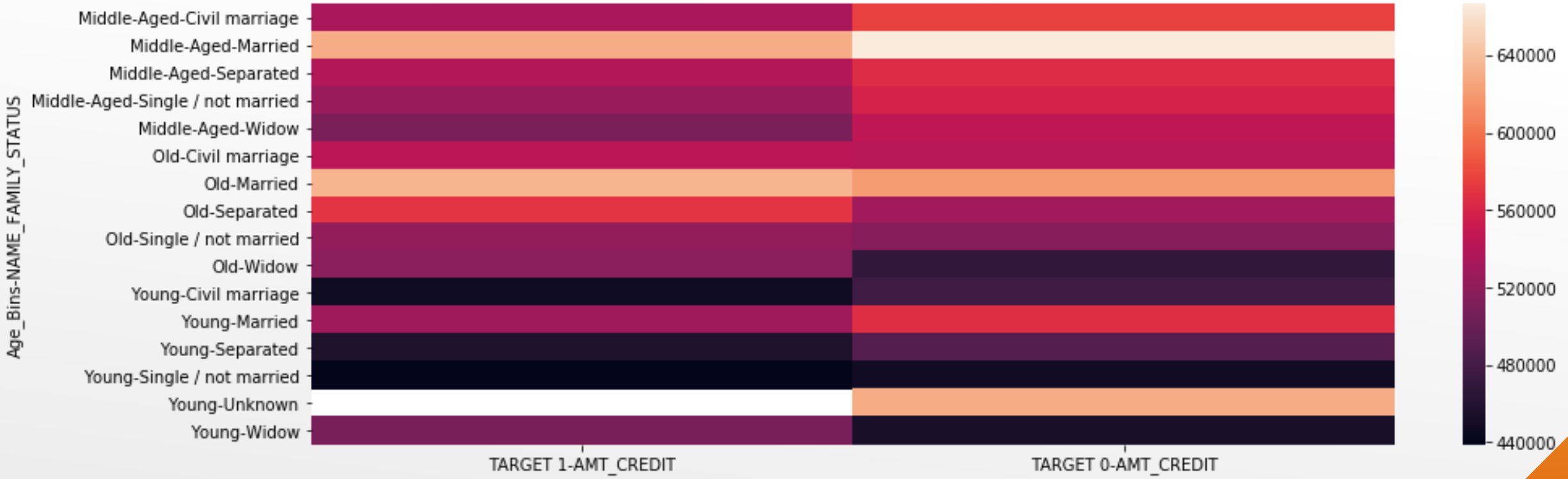


- We see young clients are more from **TARGET 1** irrespective of the income types than that in **TARGET 0**, considering the average number of employment days.

- That directly tells that Age Brackets matter for the client being a defaulter or not.

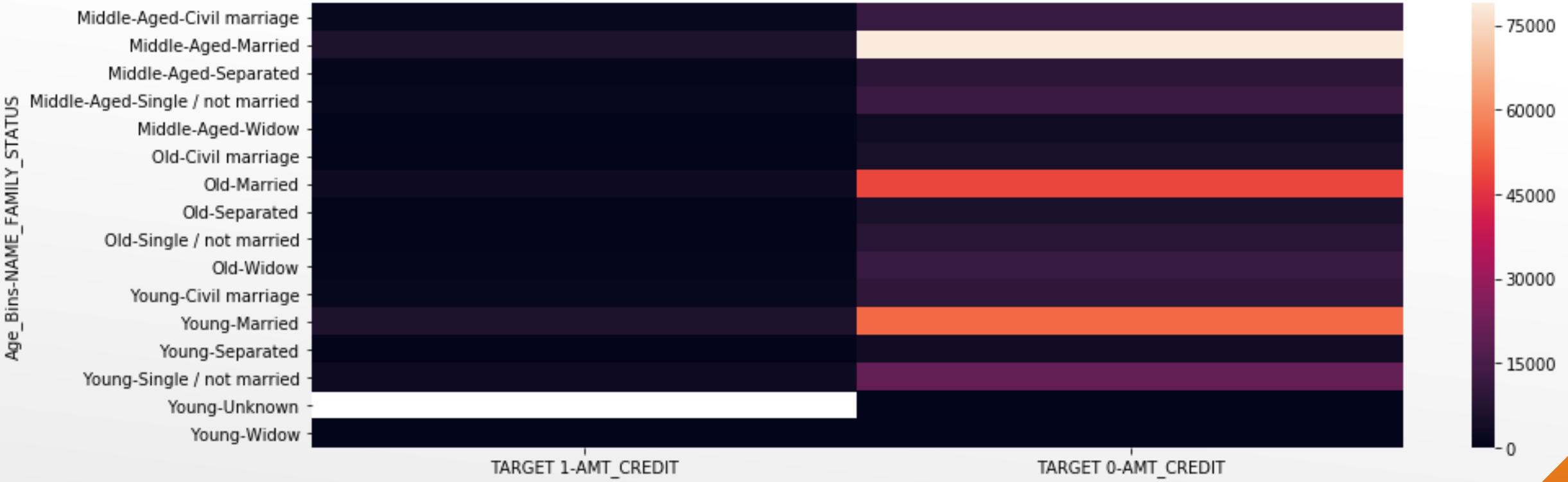
- Considering the Age brackets and type of income source, the average credit amount for **TARGET 1** is less than **TARGET 0**.

- This can be said as less orange shades lie for **TARGET 1**.

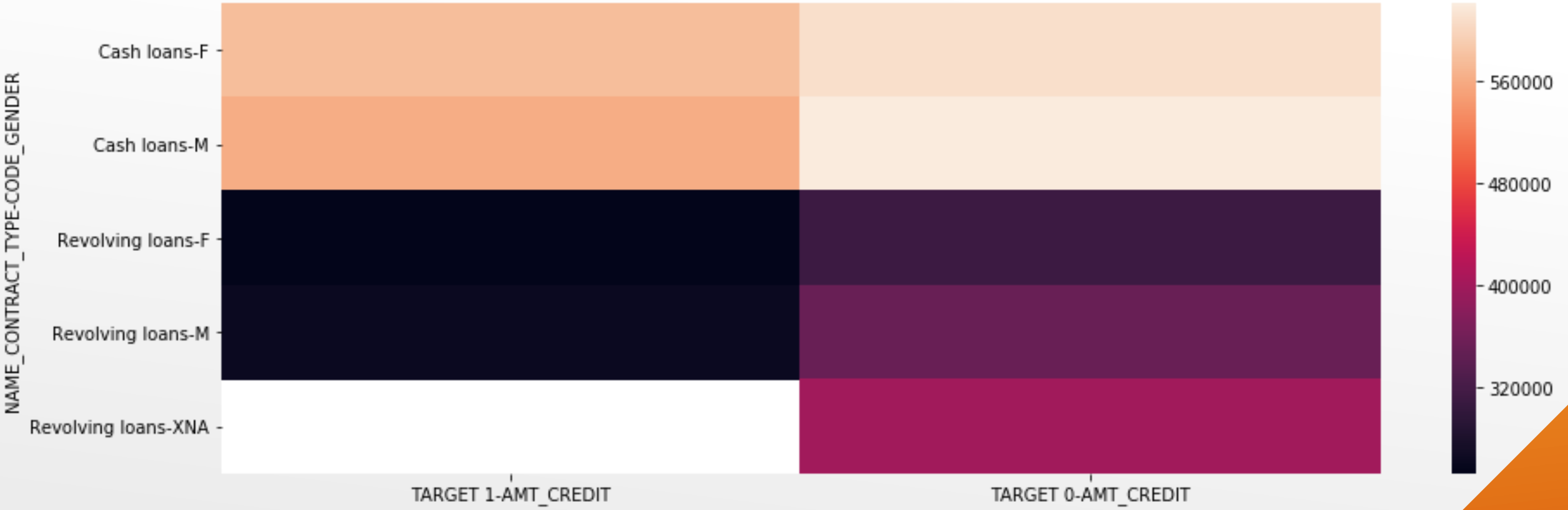


- We notice that while considering the age brackets and the family status of the clients, the average credit amount is less for **TARGET 1** clients than the **TARGET 0**.

- Mostly young generation can be seen in dark shades.



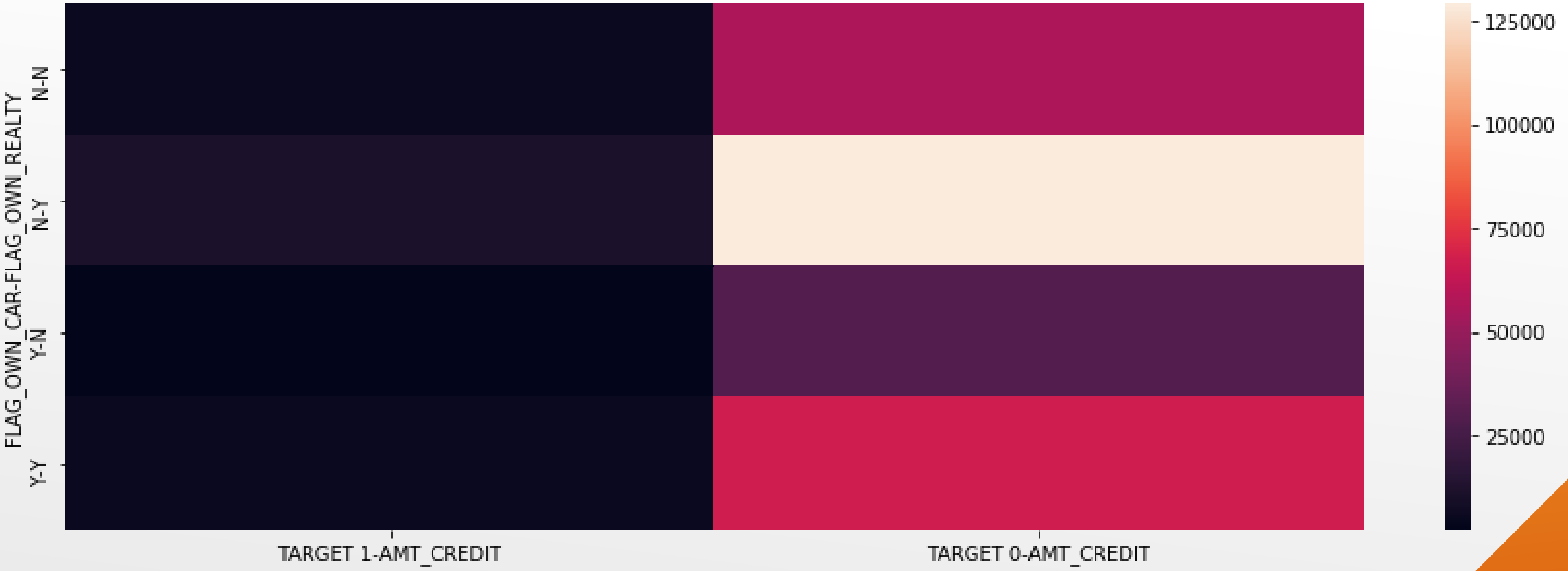
- It seems when considering the age bins and the family status, the count of loans is less for **TARGET 1** than **TARGET 0**.
- For **TARGET 0**, most loans are taken by old married people and young married people.



- Considering the type of loans (cash and revolving) and the gender of the client, we find, the average credit loan amount is way higher for **TARGET 0** clients with cash loans, while way lower for **TARGET 1** clients with revolving loans.

- Considering the type of loans (cash and revolving) and the gender of the client, we find, the number of credit loans is way lower for **TARGET 1** clients.

- When considered the owing of car and house, we find the average credit amount is way lower for those not holding cars for **TARGET 1**.
- While for **TARGET 0**, those holding cars irrespective of the house, have the high average credit amount.



- When considered the owing of car and house, we find the count of loans is way lower for **TARGET 1** irrespective of having car or house.

CONCLUSIONS

- Age_Bins (which is derived from DAYS_BIRTH) is a strong indicator of knowing whether the client will face the difficult in repaying the loan or not. We have concluded that Young people ought to be more of facing difficulties in repaying; while Middle-aged and Old's not.

- Other such indicator we found is the Income_Bins (derived from AMT_INCOME_TOTAL). The High income crowd seems to fall less in facing difficulties while repaying the loan amount.

- Another one is the WEEKDAY_APPR_PROCESS_START. More clients who are at ease of repaying the loans come for loans on Mondays than on Fridays. That is a clear-cut hint if a client comes on Fridays, he/she might be the one who can face difficulty in repaying back to the bank.

- Then comes the CNT_CHILDREN. If a client has 2 children there is more probability of him repaying back at ease, in comparison to the ones who bear no child.

- NAME_FAMILY_STATUS is also important since we noticed that Widows are more liable of facing difficulties in repaying back.

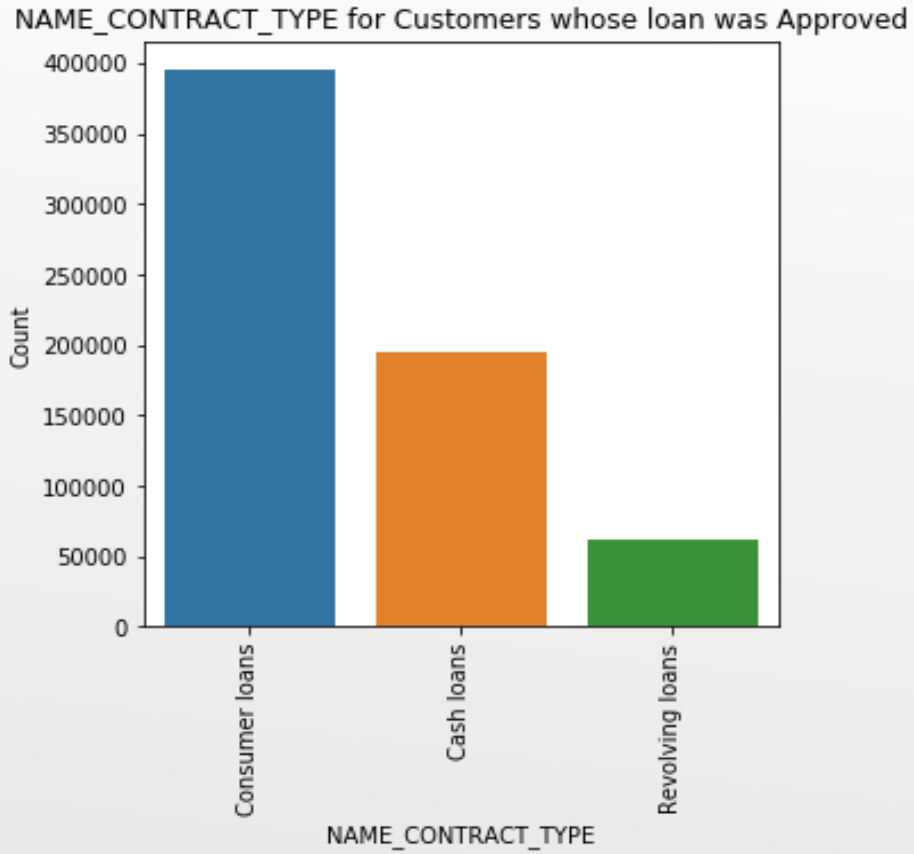
- Considering NAME_INCOME_TYPE, we see loans can be given to Businessman and Students with low risk rates, since they do not fail in repaying back, while giving loans to people on maternity leaves is risky.

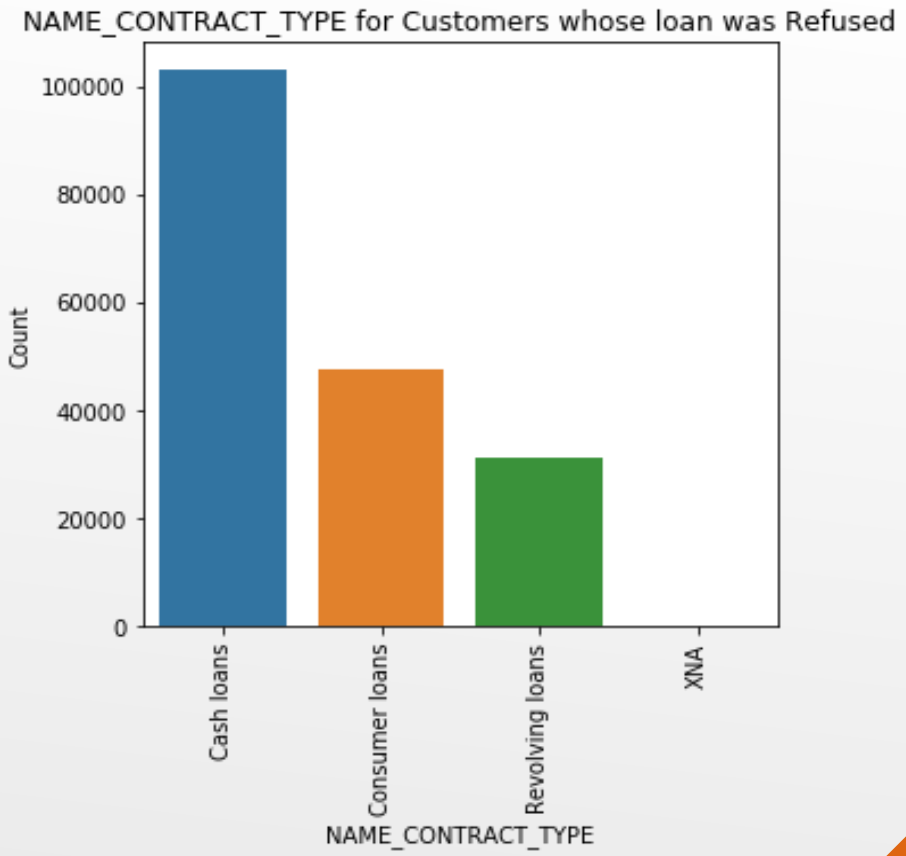
- DAYS_EMPLOYED says that clients with less number of employment days are more inclined towards facing difficulties in paying them back.

PREVIOUS DATA EDA

- We will now check the 2nd file which contains information about the client's previous loan data.
- It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

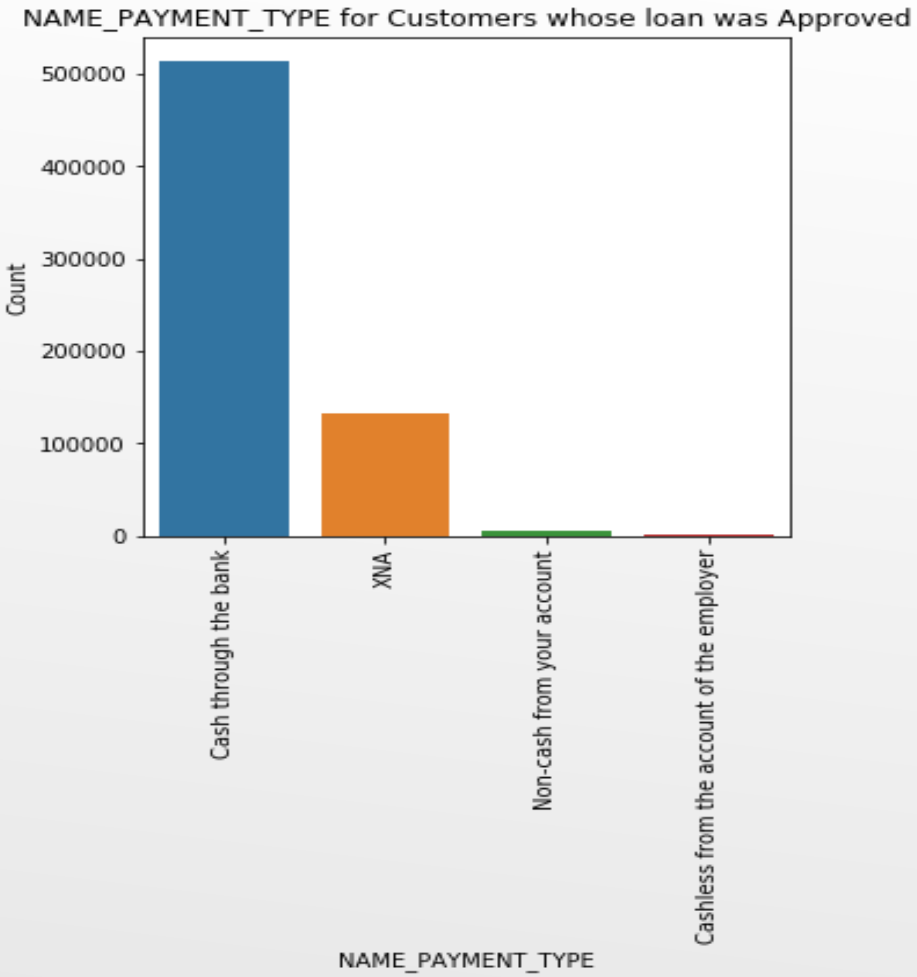
PLOTS AND INFERENCES

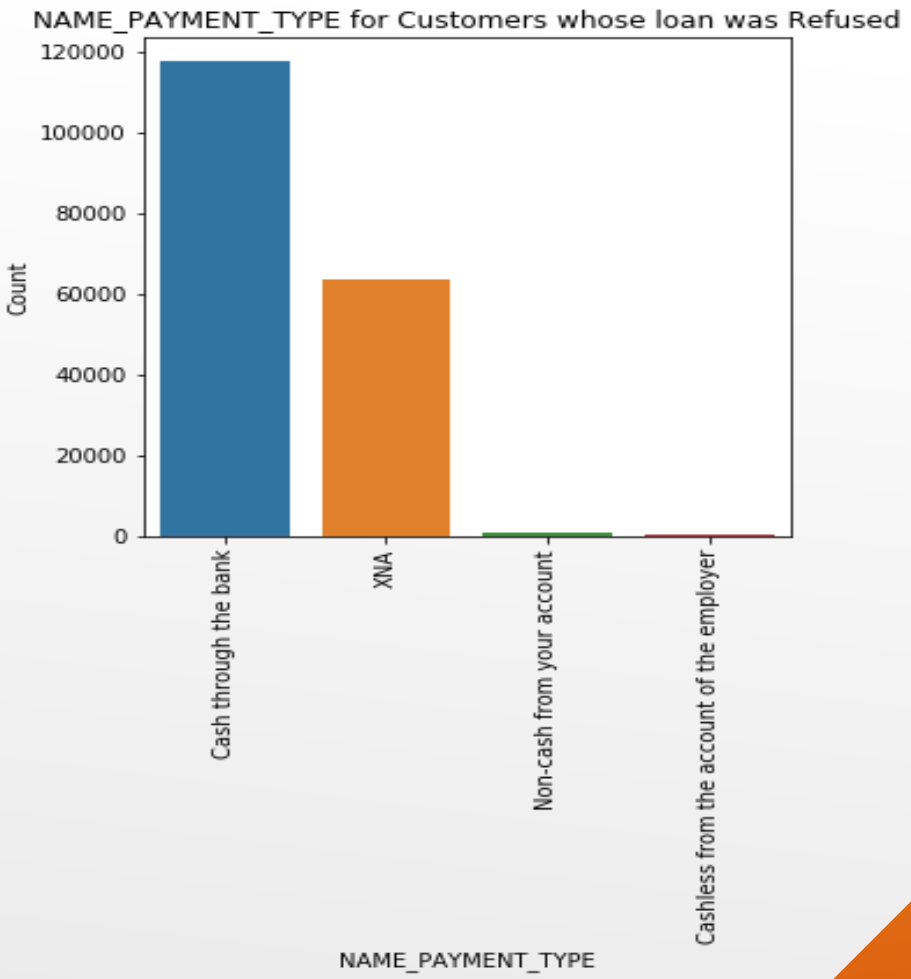




- For Loans which were **Approved**, the Consumer Loans were more than others and for **Refused** loans, the count of Cash Loans were more.

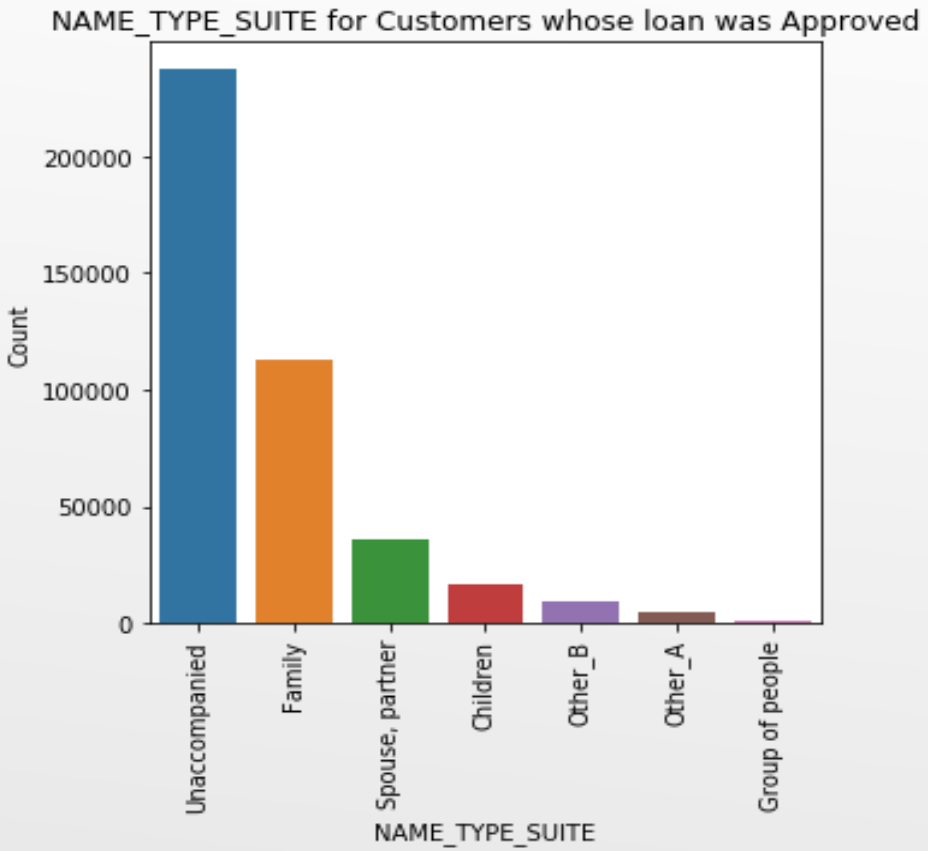
- This means that the **Approval** rate for Consumer loans is more than Cash and Revolving loans.

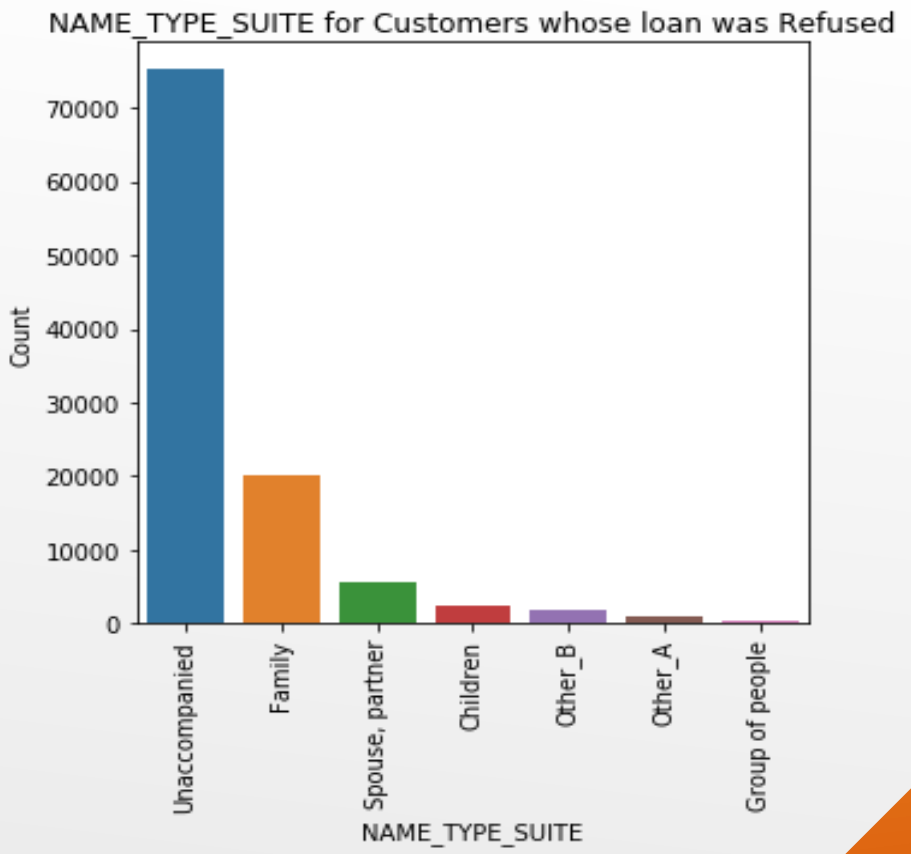




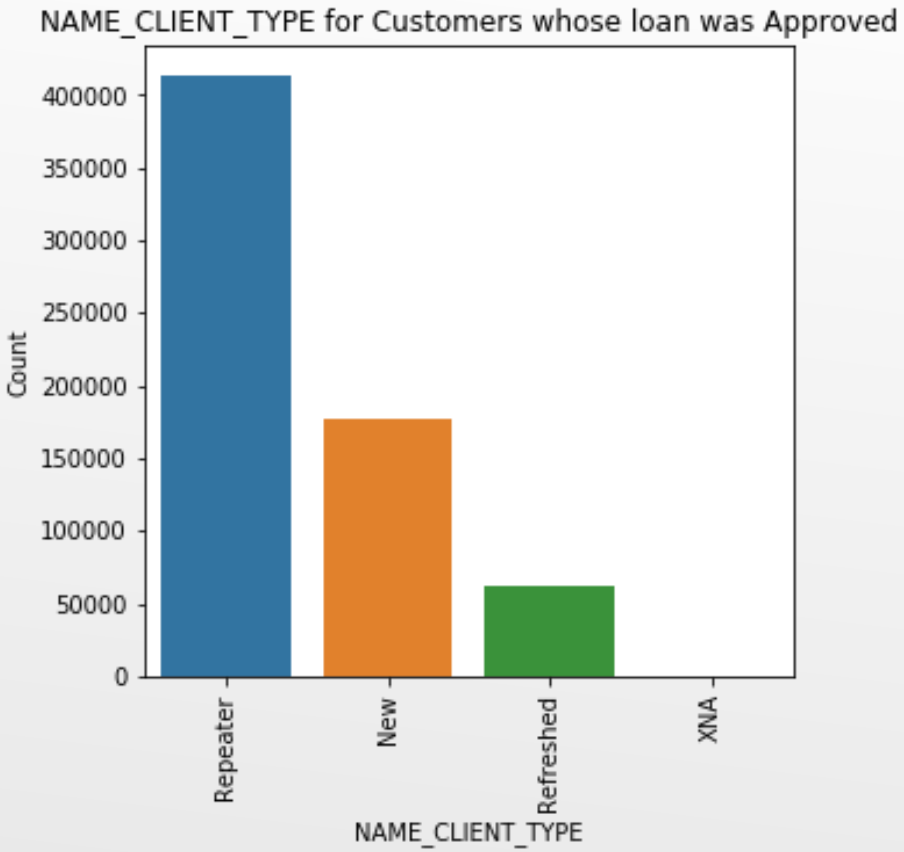
- In both outcomes of loan status, we can see that the users preferred to pay via the Cash through the bank option as opposed to other options provided to the customer.

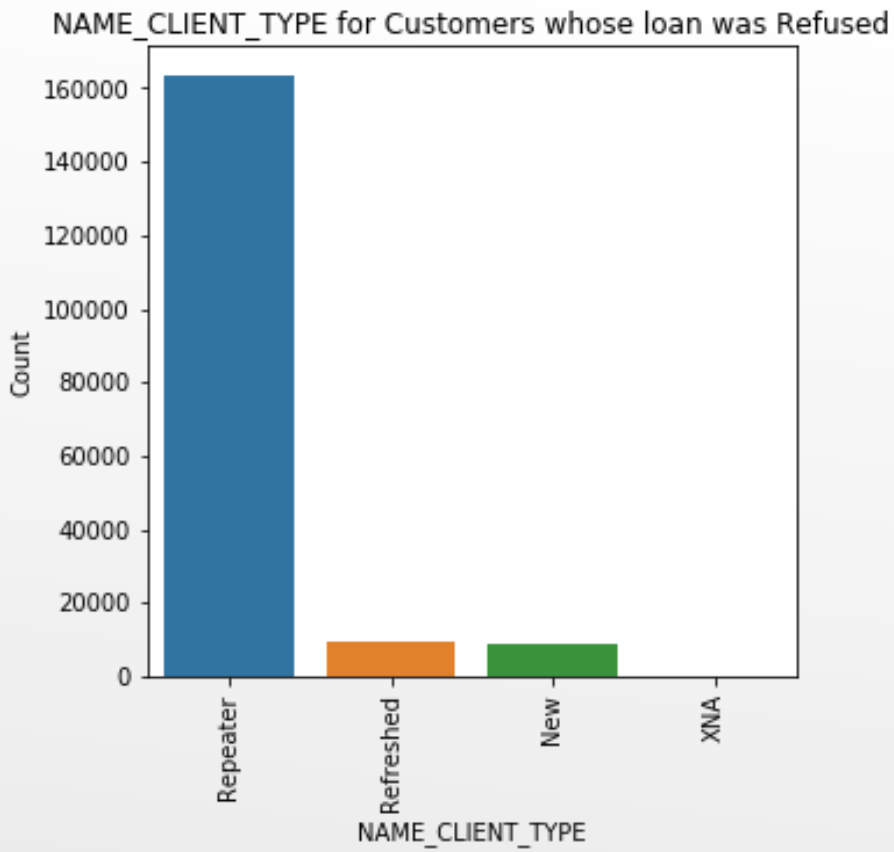
- **Note**: we are treating XNA value as not available hence please ignore the same.



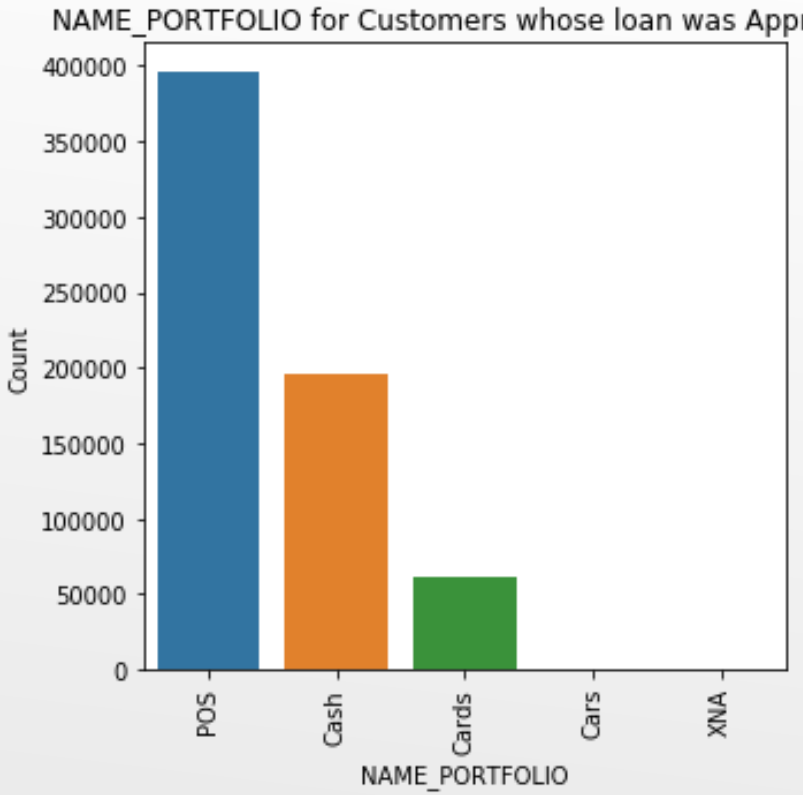


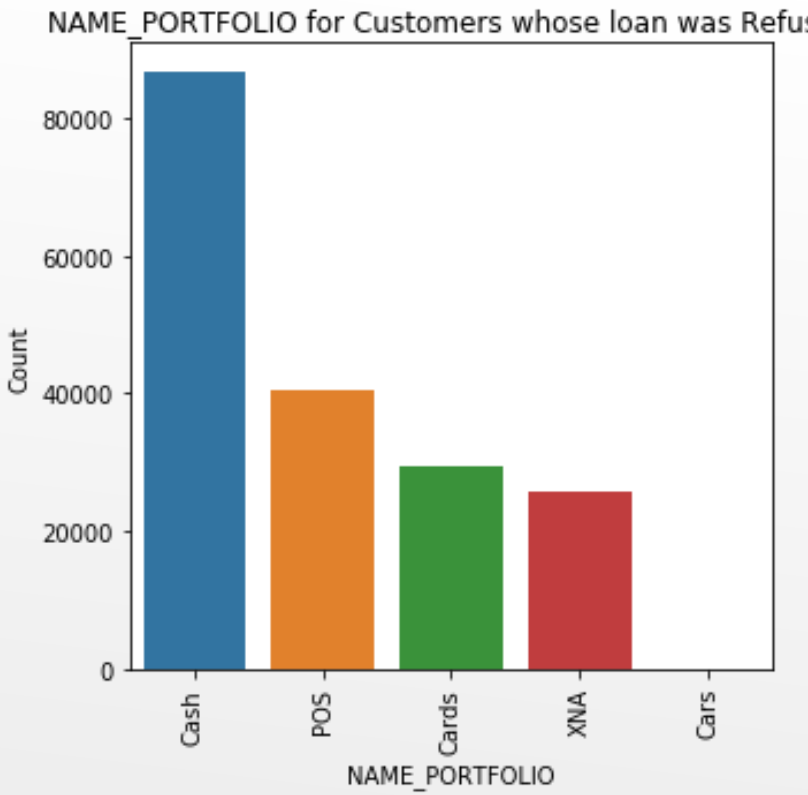
- Most of the clients who apply for loans are Unaccompanied. And there is a drastic drop to the ones who are accompanied with their Spouse **and** children or a Group of People.
- The **Approval** rate for Users accompanied with Family or who are **Unaccompanied** is more than others.



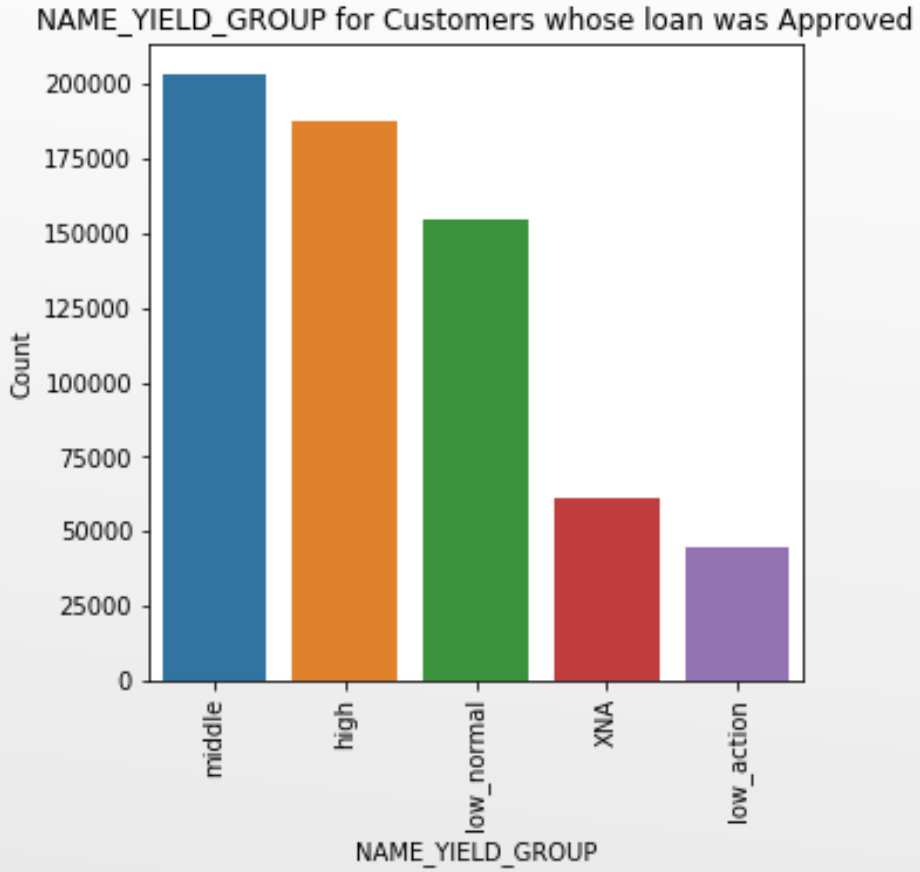


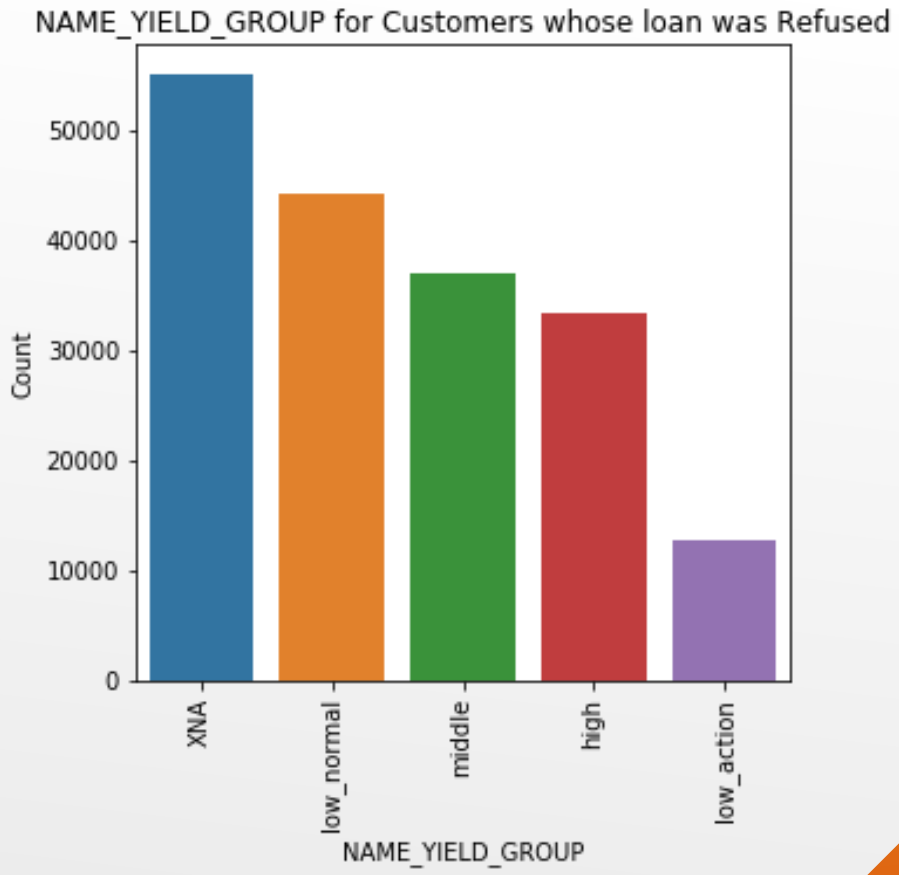
- Majority of the customer who apply for a loan are Repeat customers.
- If the customer is a New customer, then the probability of them having their Loan **Approved** is more than **Refusal** of the loan.



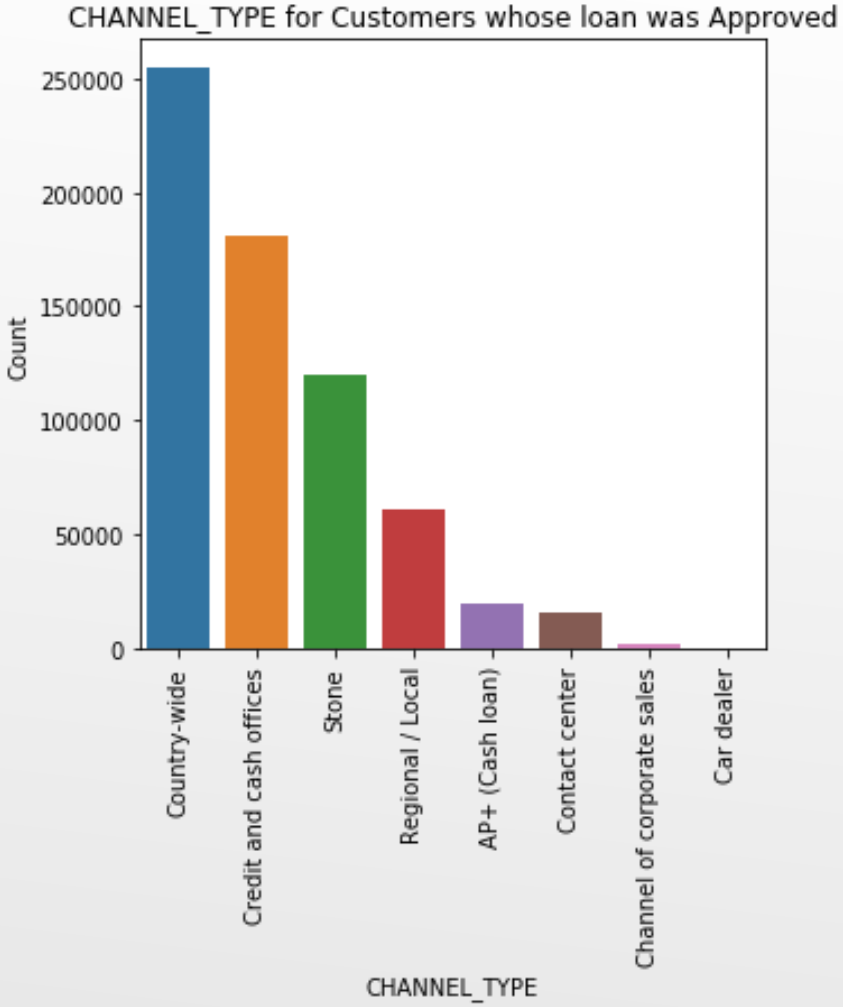


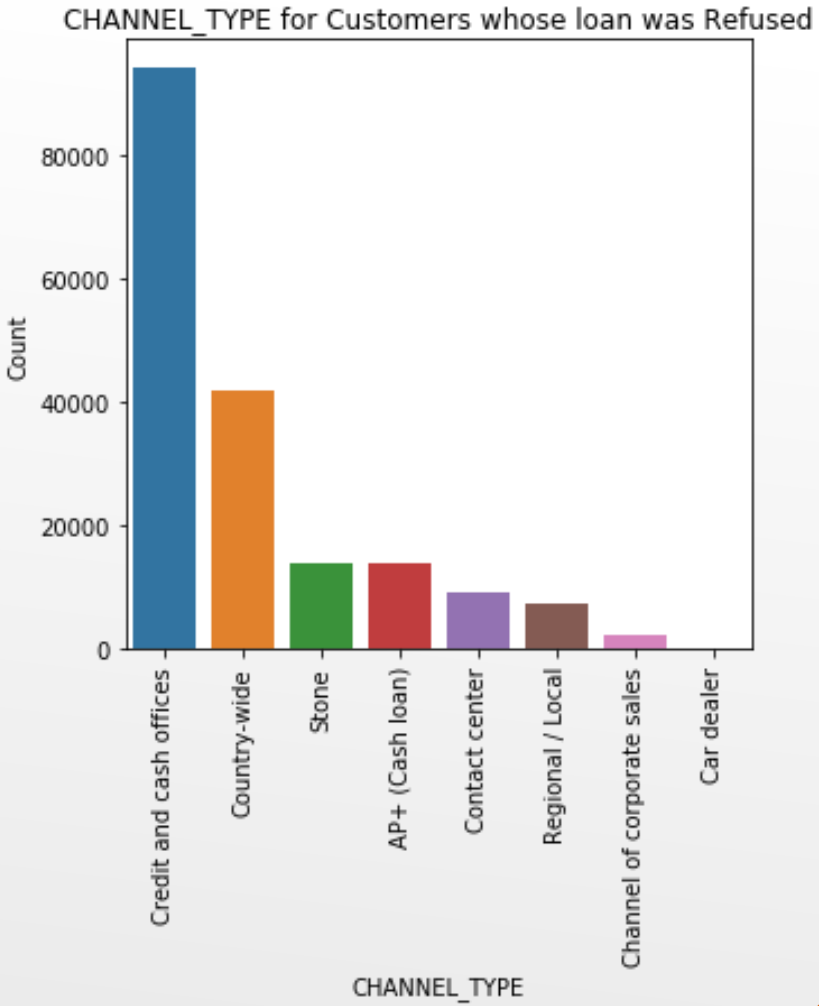
- For Loans which were **Approved**, we can see that the application was more for POS as opposed to Cash for loan **Refusals**.
- Since the **Approval** rate for customers wanting POS is more than any other type of loan type, we should target users who go for POS option.
- Since the **Refusal** rate for customers wanting Cash is more than any other type of loan type, we should avoid users who go for Cash option.





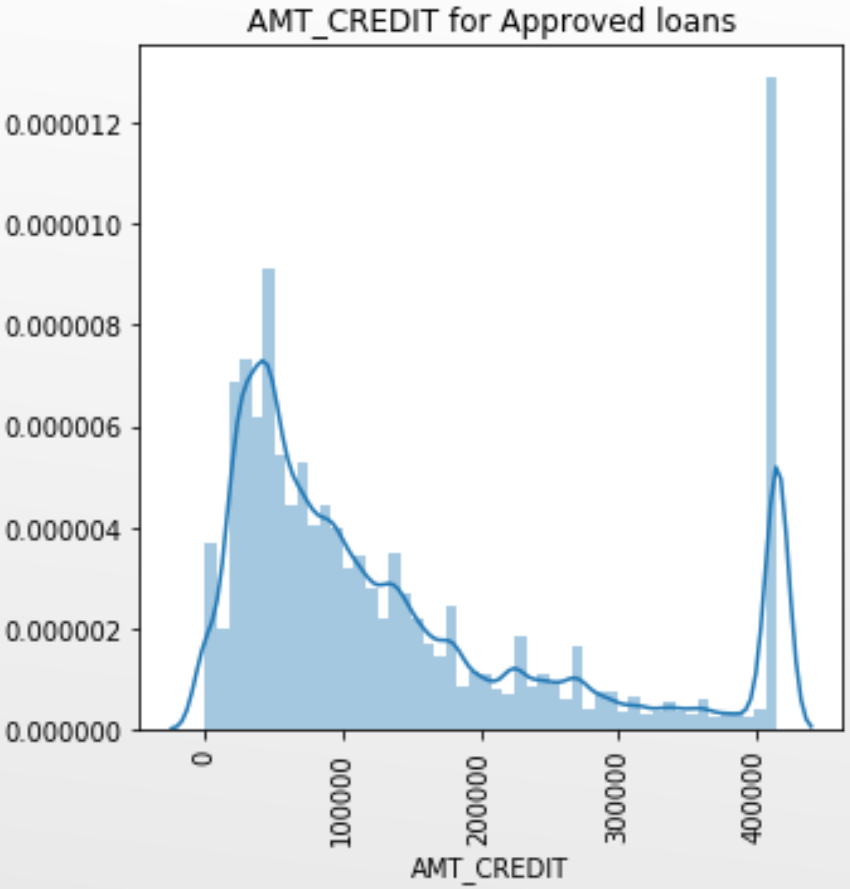
- We can also see that if the rate of interest is in the middle group then the Loan **Approval** is more as opposed to Loan **Refusals** which is high for low_normal interest rates.

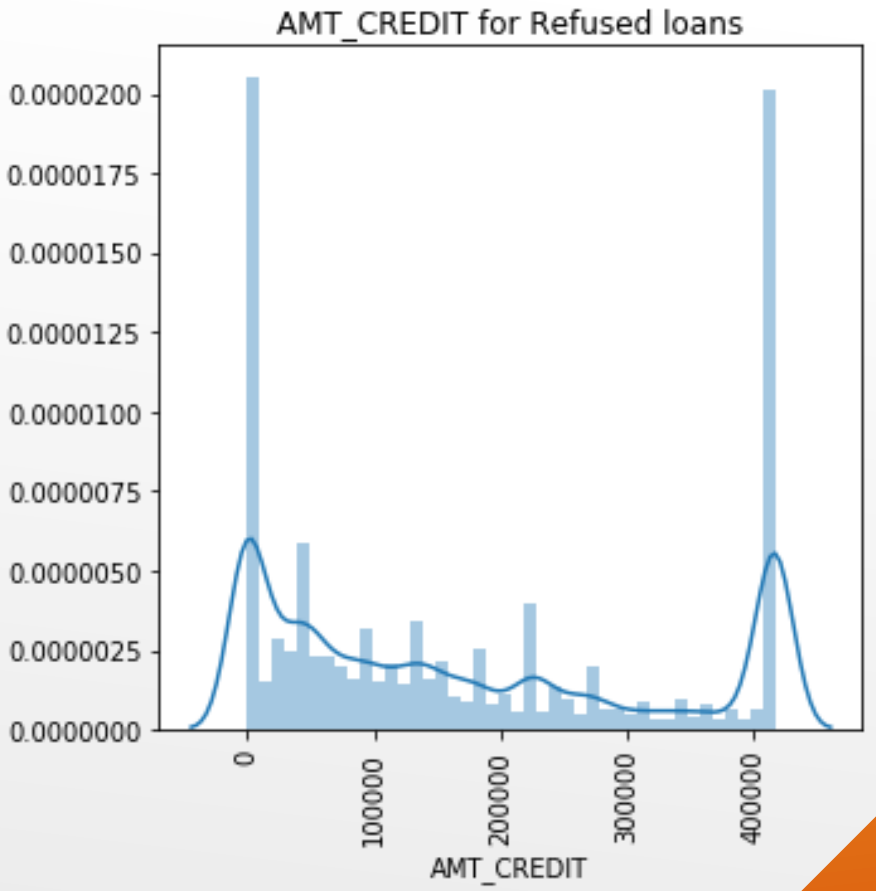




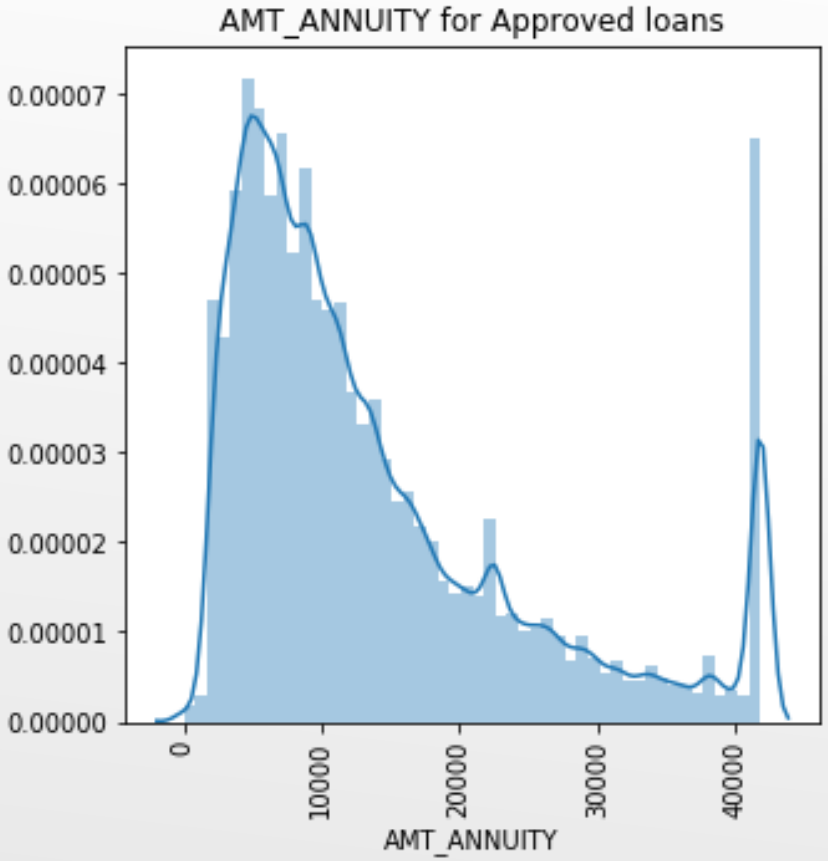
- We can also see that we should prioritize the loan application for clients from Country-wide search since these clients have a higher rate of Loan **Approvals**.

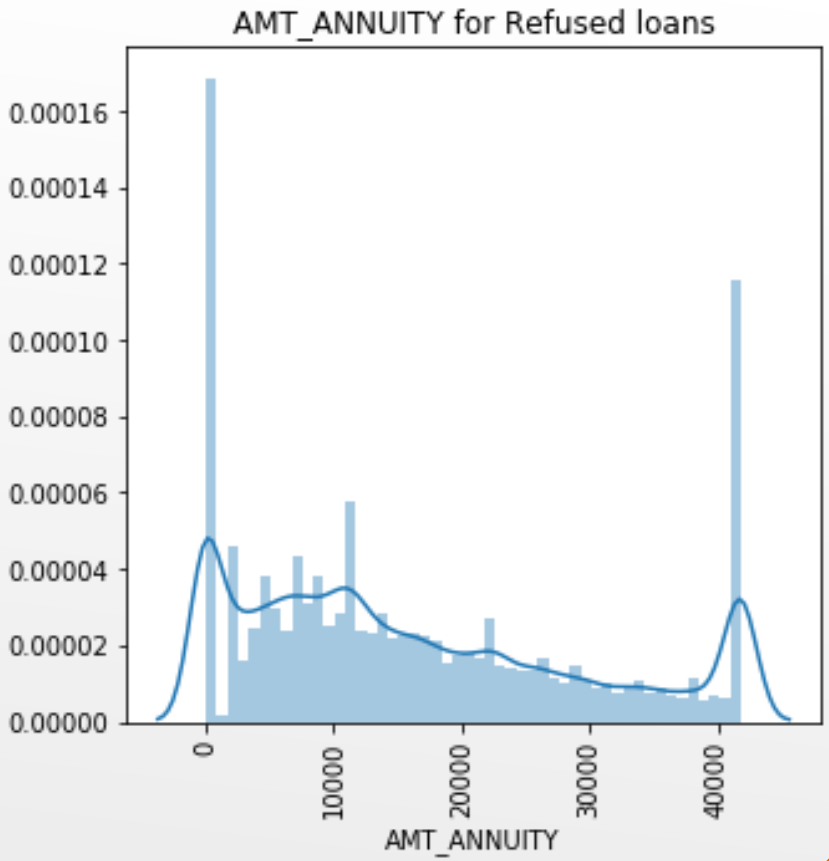
- We should also avoid/ deprioritize the loan application for clients who are from Credit and cash offices since these clients have a higher rate of Loan **Refusals**.



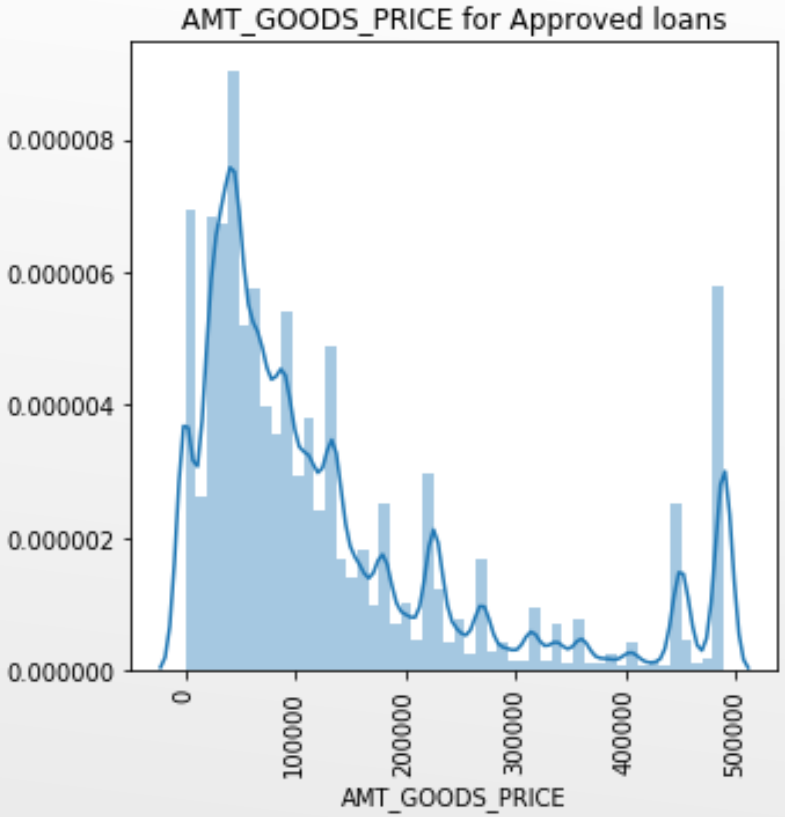


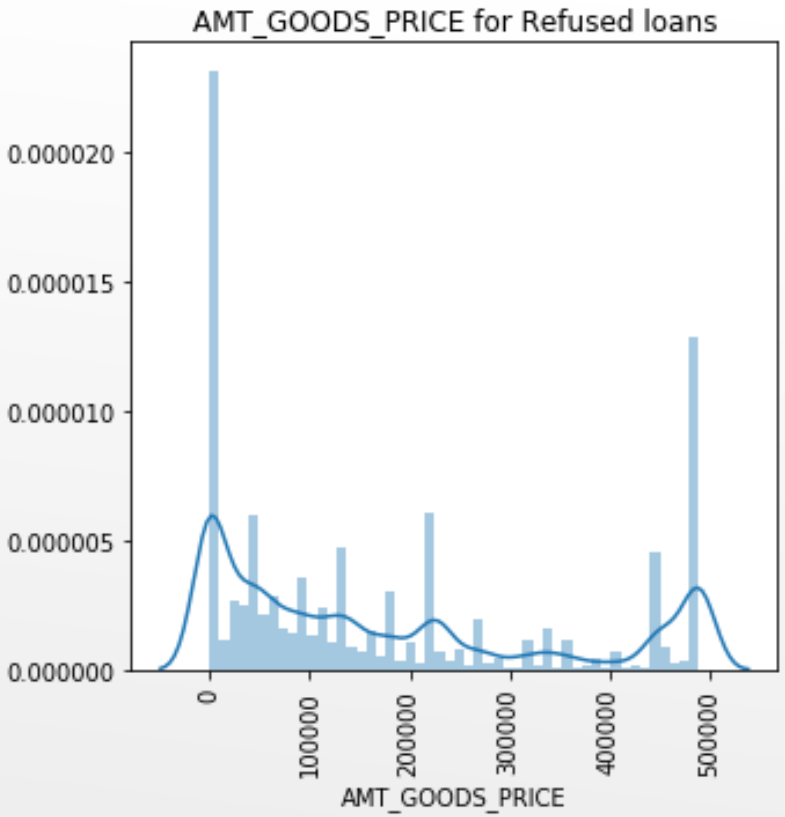
- We can see clients who had **Approved** loans opt for loans amount between **50,000 and 1,00,000.**
- For **Refused** loans, there is a spike at amount of approx. **10000**.
- Also an interesting thing is that there is a sudden spike at **4,00,000** for both categories.



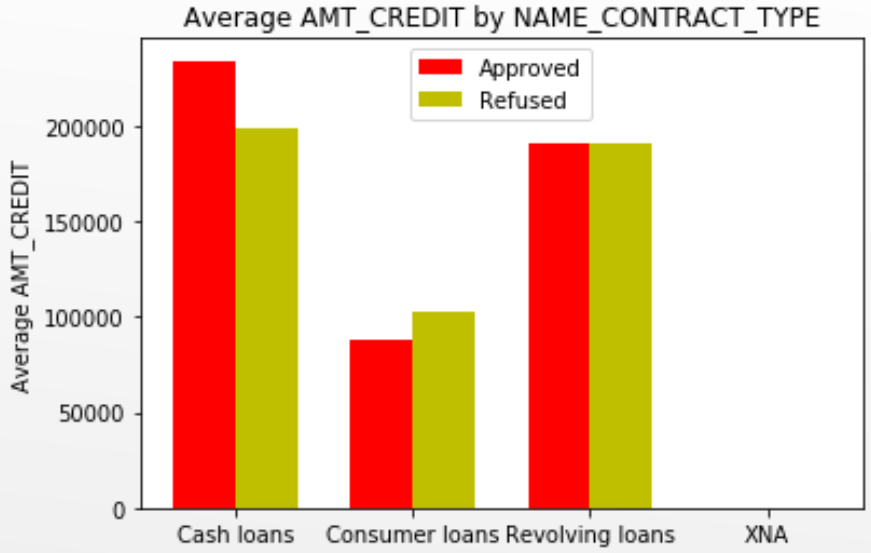


- As seen in loan credit amount, there is a similar spike in **annuity** amount at **42,000**.

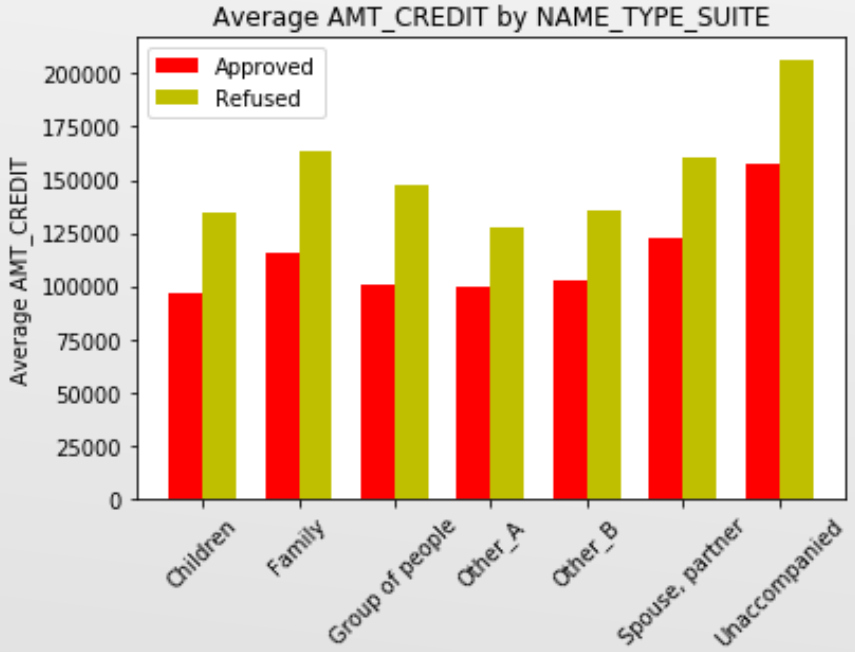




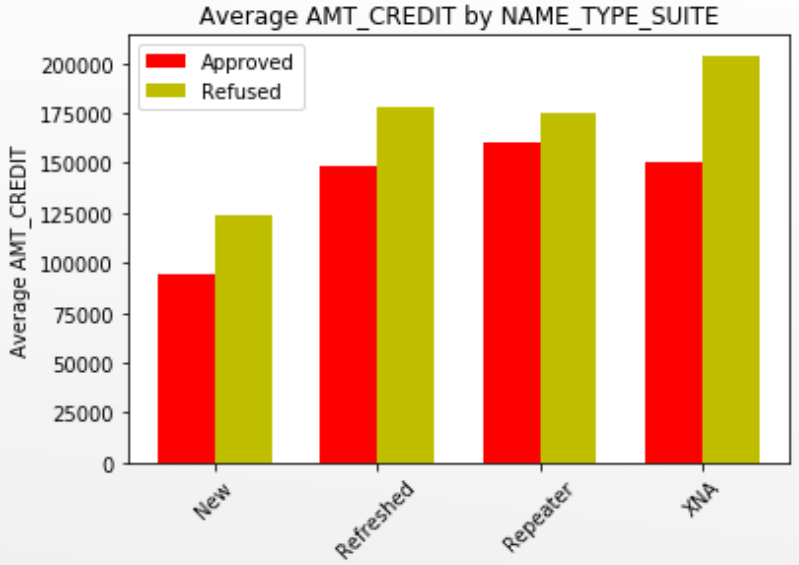
- Clients whose loan was **Approved** have more loans for goods amounting of **50,000** in comparison to **3,00,000 and 4,00,000** (where it is the lowest) while those whose loan was **Refused** have goods amounting of either **10,000 or 4,80,000** in value.



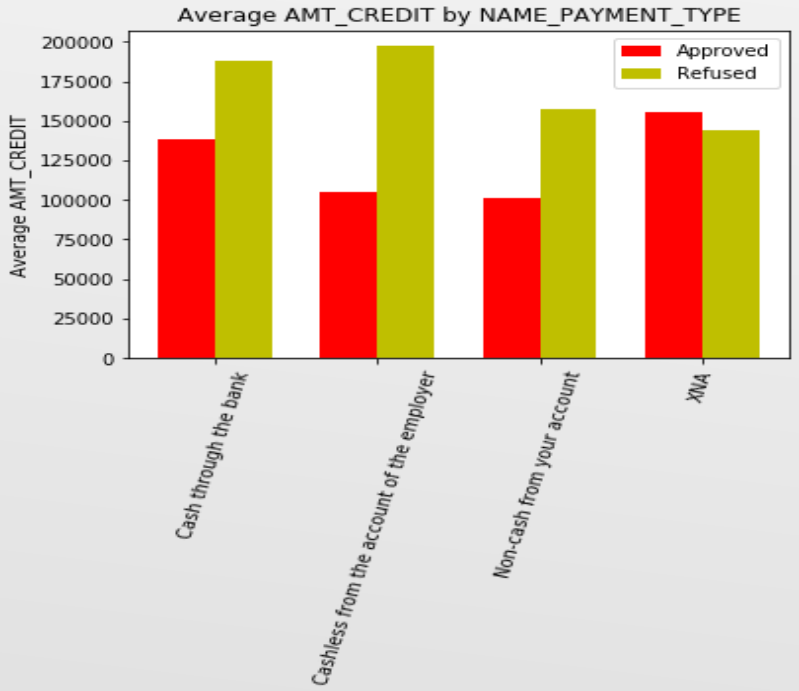
- The Average Credit Amount for **Approved** Loans is more for **Cash** loans and least for **Consumer** loans but it is the opposite for **Refused** Loans.
- The Average Credit Amount for **Revolving** Loan type is same for both **Approved** and **Refused** loans.



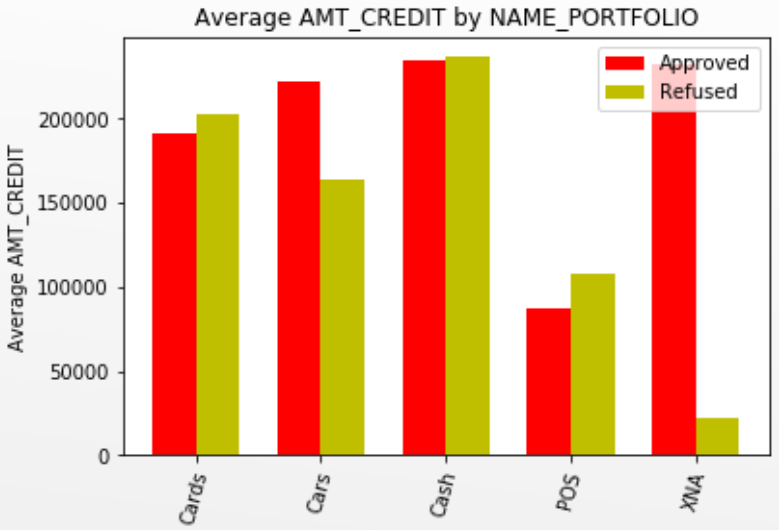
- The average credit amount for any client accomplice have been a higher for **Refused** Loans than **Approved** loans.
- The average credit limit is more for clients who came **Unaccompanied** to the bank as opposed to Children or a Family or with a **Group of people** for Approved loans.
- The average credit limit is more for clients who came **Unaccompanied** to the bank as opposed to **Other_A or Other_B** category for Refused loans.



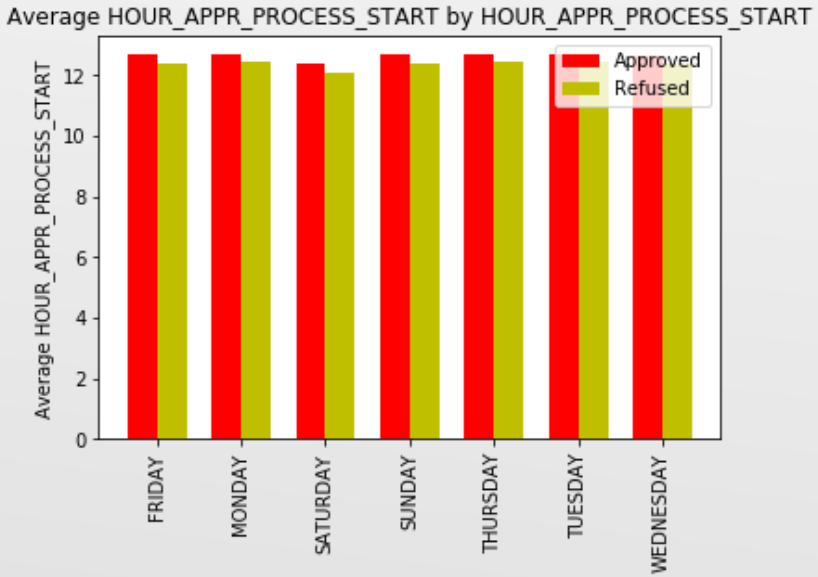
- The average credit amount for any type of client(**New/Repeater/Refreshed**) has been higher for **Refused** Loans than **Approved** loans.
- The average credit amount is more for **Repeater** Clients as opposed to **New** Clients for **Approved** loans.
- The average credit amount is more for **Refreshed** Clients as opposed to **New** Clients for **Refused** loans.



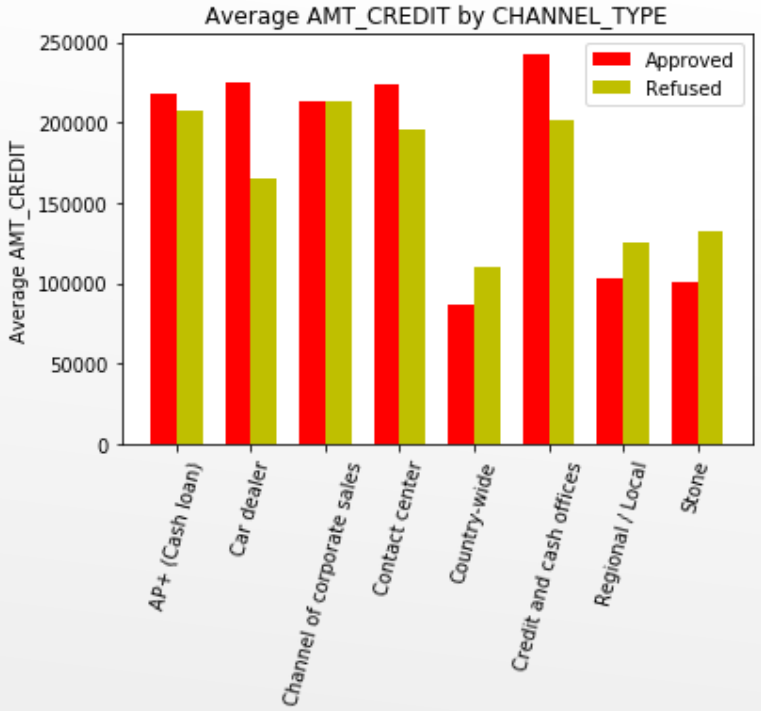
- The average credit amount for any kind of payment type have been higher for **Refused** Loans than **Approved** loans.
- The average of the credit amount is more for **Cash through the bank** payment type as opposed to **Non-Cash from your account** payment type for **Approved** Loans.
- The average of the credit amount is more for **Cashless from the account of the employer** payment type as opposed to **Non-Cash from your account** payment type for **Refused** Loans.



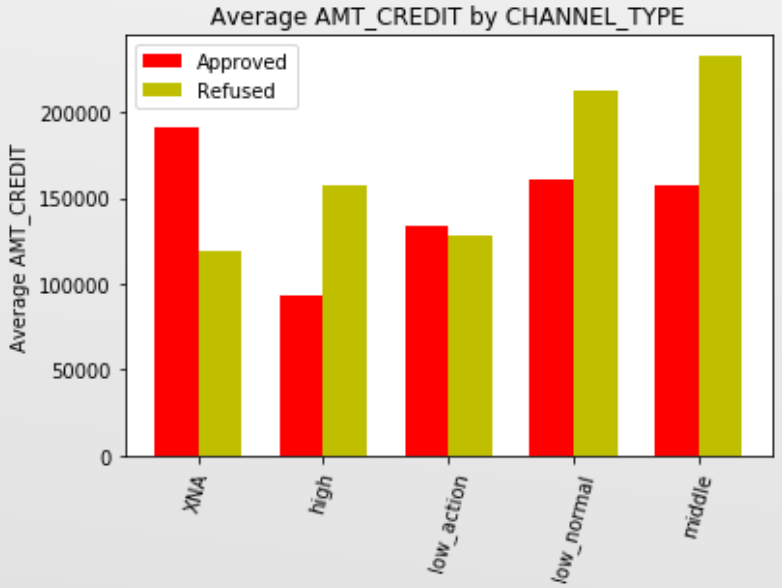
- The average credit amount for any type of portfolio has been slightly higher for **Approved** Loans than **Refused** loans.
- The average credit amount has been more for clients who were looking to buy a **Car** as opposed to **POS** or **card** payment for **Approved** loans.
- The average credit amount has been more for clients who took a **Cash loan** as opposed to **Car loan** or POS for **Refused** loans.



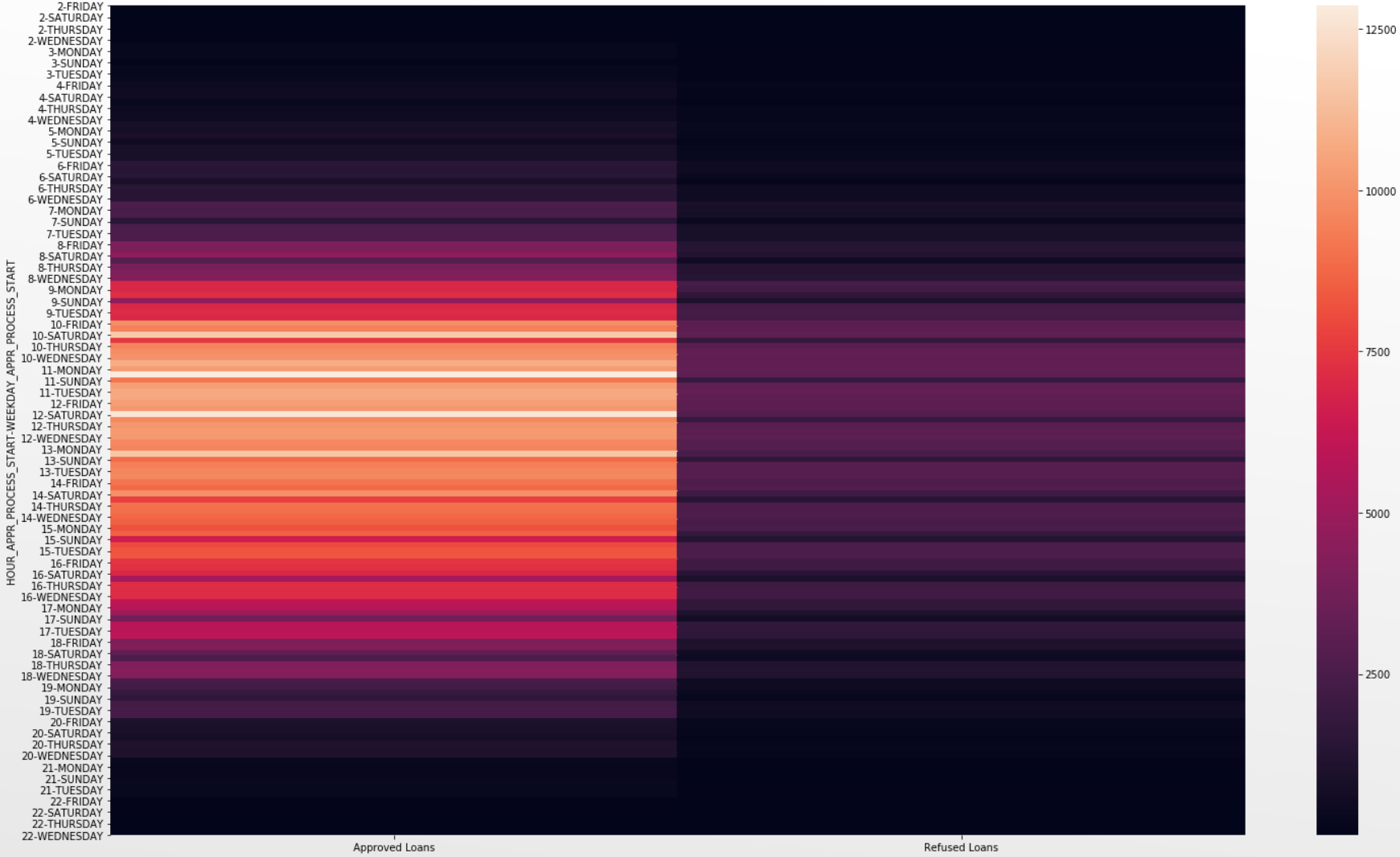
- Loan application starting process is generally seen at 12 PM in the afternoon, let it be any day.
- To be precise, the clients whose loan is **Approved** come slightly later on any day for loan applying than clients whose loan is **Refused**.



- The average credit amount has been more for **Approved** loans as compared with **Refused** loans.
- The average credit amount is more for **Credit and Cash Offices** and least for **Country-wide** channel for **Approved** loans.
- The average credit amount has been more for **Channel of corporate sales** as opposed to **Country-wide** channel for **Refused** loans.

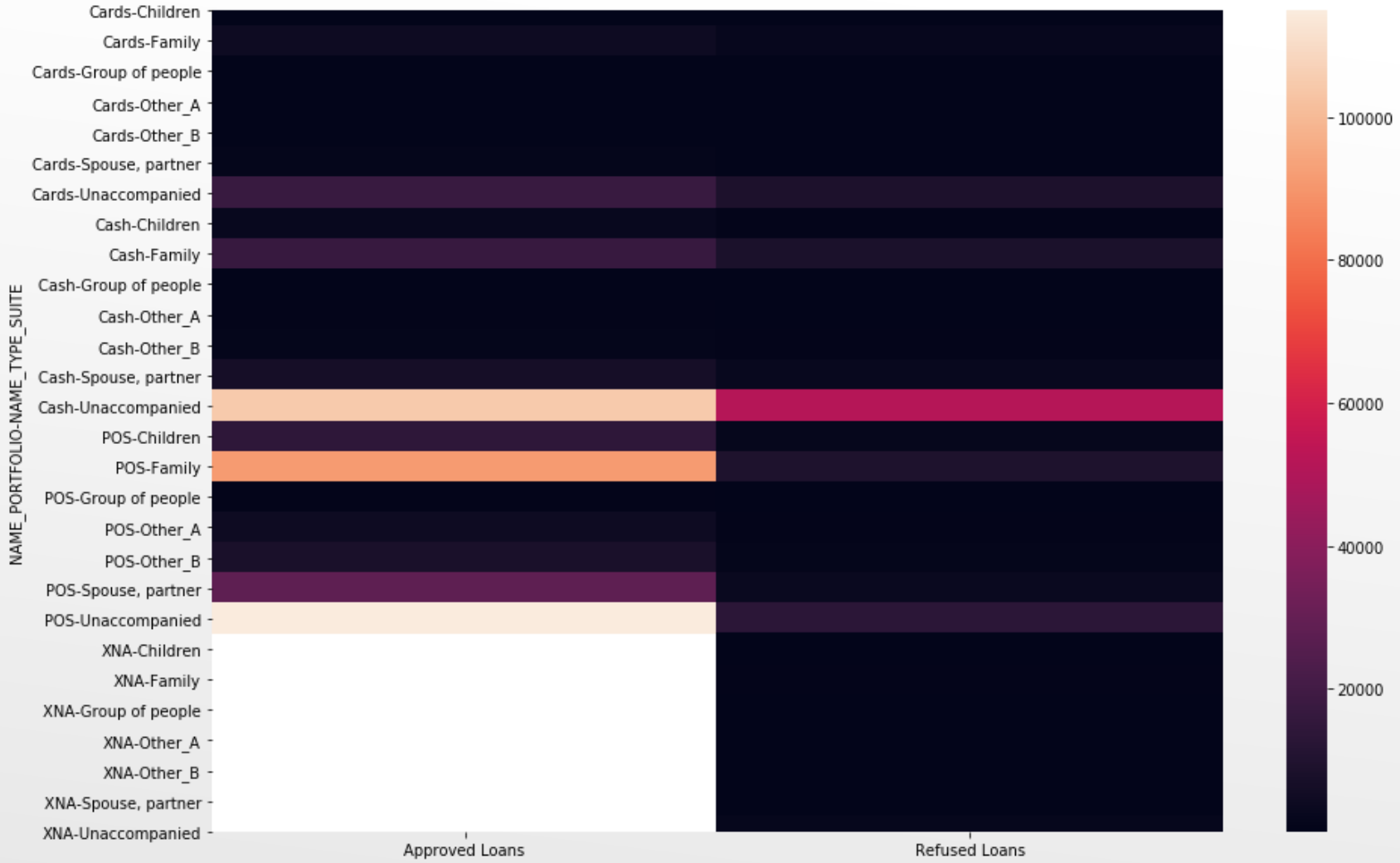


- The average credit amount has been more for **Refused** loans as compared with **Approved** loans for different interest rates.
- The average credit amount is more for **low_normal** interest group and least for **high** interest group for **Approved** loans.
- The average credit amount has been more for **middle** interest group as opposed to **low_action** interest group for **Refused** loans.



- As we can see that the most amount of **Approvals** are done when the application is started on Saturday from **10 AM to 1 PM** or if it is better to start the application for the clients between **9 AM till 4 PM** on any given day.

- We can also see that the most amount of **Refusals** are done when the application is started either before **8 AM or after 6 PM** on any given day.

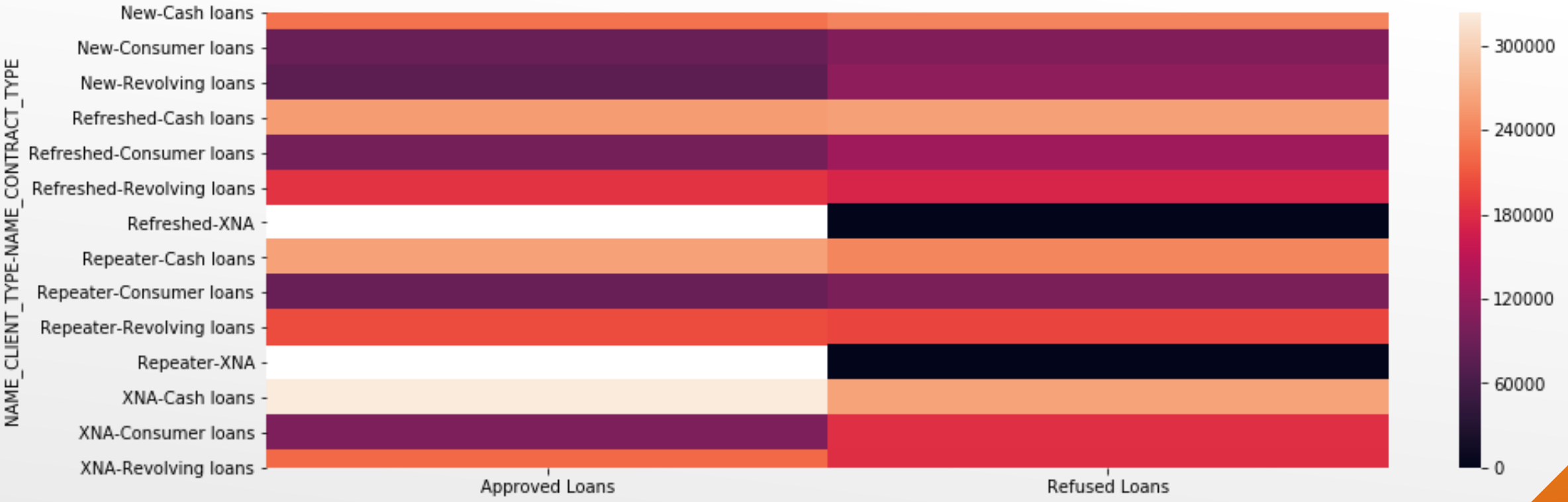


- As we can see, if the client is asking for **Cash** loan and who had come **Unaccompanied** to the bank or if a client asks for a **POS** loan who came in with a **Family** stands a better chance for loan **Approval**.

- For **Card** loan, both the **Approval** and **Refusal** count is low.

- As we can see, if a **New client** is asking for **Consumer loan** or a **Repeater client** asking for **Cash loans** stands a better chance for loan **Approval**.

- But if a **Repeat client** asks for either a **Consumer** or a **Revolving** Loan, then the **Refusal** rate is more as compared to others.



- As we can see, if a **New client** is asking for **Cash loan** and the average loan amount is around **2,30,000** then the loan **Approval** rate is higher but if the same exceeds **2,40,000** then the **Refusal** rate increases.

- But if a **Repeat or a Refreshed customer** asks for a **Consumer Loan** and the average loan amount is below **1,00,000** then we have a better chance of **Approvals** but if the same amount is more than **1,00,000,** then the **Refusal** rate increases.

- As we can see, if a **New client** is asking for **Cash loan** and the average annuity is around **21,000** then the loan **Approval** rate is higher but if the same exceeds **25,000** then the **Refusal** rate increases.

- But if a **Repeat or a Refreshed customer** asks for a **Consumer Loan** and the average annuity amount is below **10,000** then we have a better chance of **Approvals** but if the same amount is more than **10,000,** then the **Refusal** rate increases.

CONCLUSIONS

Below factors contribute to loan **Approvals**:

1. The applications should be started on either **Saturday** from **10 am to 1 PM** or start the application for the clients **between 9 AM till 4 PM on any given day**.
2. If the client is asking for **Cash loan** and who had come **Unaccompanied** to the bank or if a client asks for a **POS** loan who came in with a **Family**.
3. If a New client is asking for **Consumer loan** or a **Repeater client** asking for **Cash loans** stands a **better** chance for loan **Approval**.
4. If a **New client** is asking for **Cash loan** and the average loan amount is **less** than equal to **2,30,000**.
5. If a **Repeat or a Refreshed customer** asks for a **Consumer Loan** and the average loan amount is below **1,00,000**.
6. If a **New client** is asking for **Cash loan** and the average **annuity** is **less than 21,000**.
7. If a **Repeat or a Refreshed customer** asks for a Consumer Loan and the average **annuity** is **less than 10,000**.

Below factors contribute to loan **Refusals**:

1. When the application is started either **before 8 AM or after 6 PM on any given day**.
2. If a Repeat customer asks for either a **Consumer** or a **Revolving Loan**, then the **Refusal** rate is more as compared to others.
3. If a **New client** is asking for **Cash loan** and the average loan amount is **more** than **2,40,000**.
4. If a **Repeat or a Refreshed customer** asks for a **Consumer Loan** and the average **loan amount** is **above 1,00,000.**
5. If a **New client** is asking for **Cash loan** and the average **annuity** is **more than 25,000**.
6. If a **Repeat or a Refreshed customer** asks for a **Consumer Loan** and the average **annuity** is **more than 10,000**.