CREDIT EDA CASE STUDY.

- By Rohit Ram & Ameya Shukla

Problem Statement

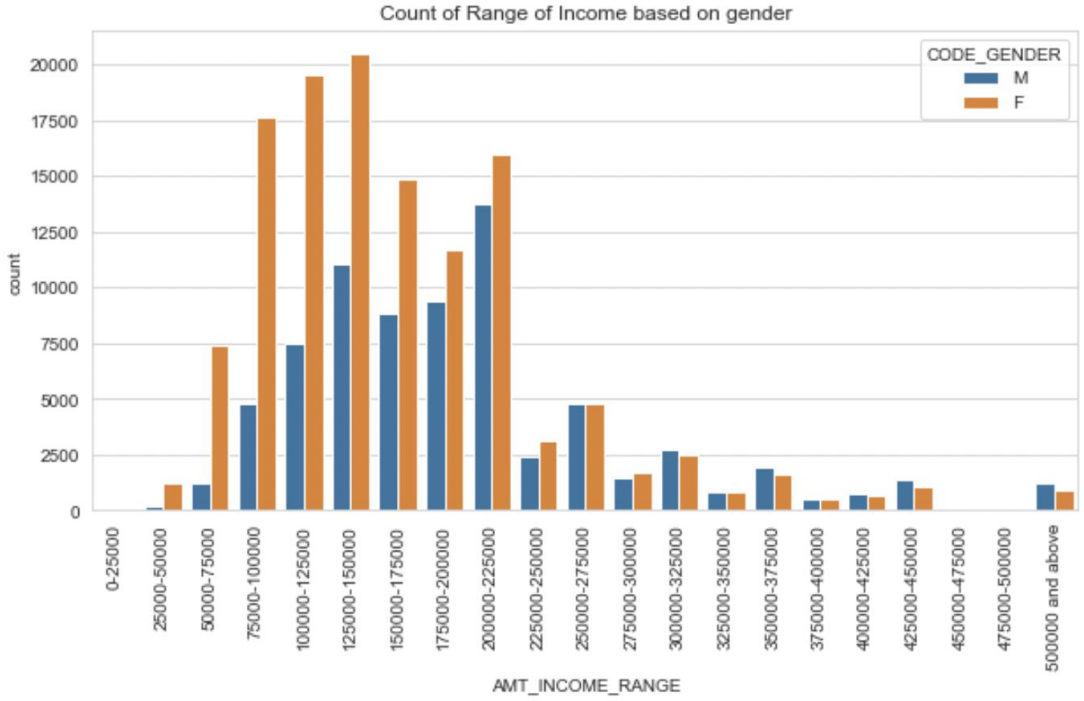
We have to use EDA to analyse the insights present in the dataset to ensure that the applicants who are capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

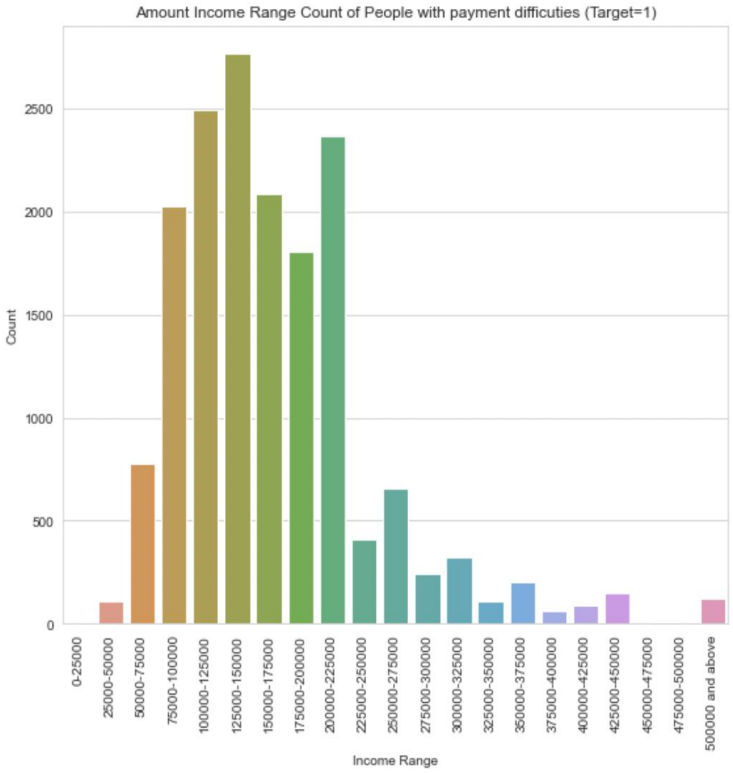
The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

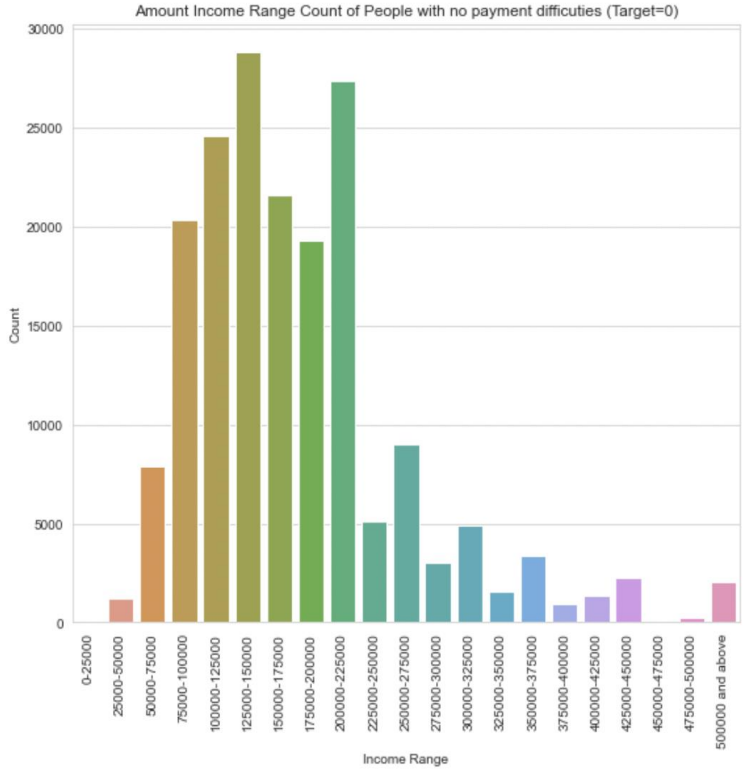
Univariate Analysis on New Application



- In the previous histplot we saw the count of Income of both the Gender, and here we have it in `Income-Range`, If we see the above countplot, Females have more number of count as compare to Males from the range `25000-225000`.

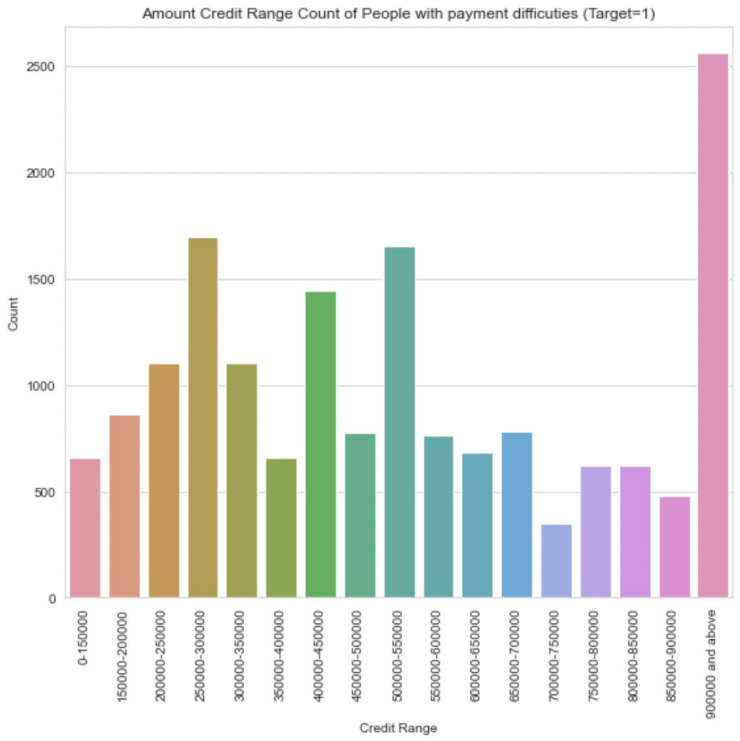
- After `225000` we can see the number of count of Males and Females are decreasing.

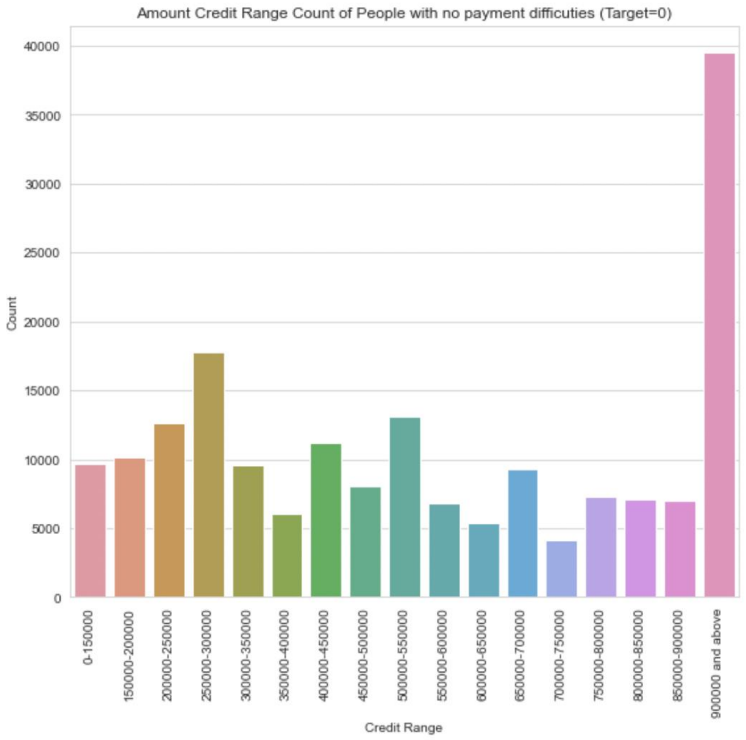




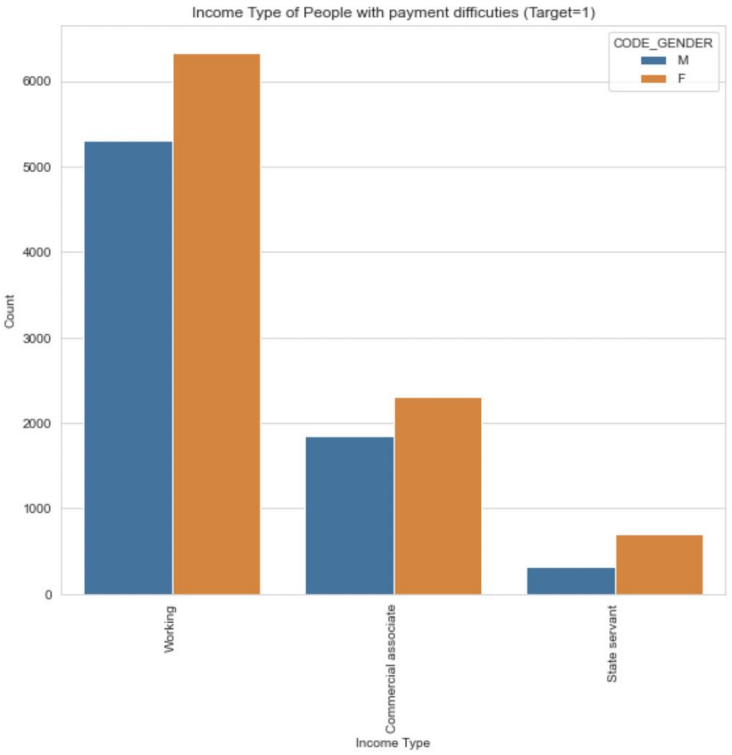
- If we see the Income-Range count with respect to `target0` and `target1`, it is almost similar.

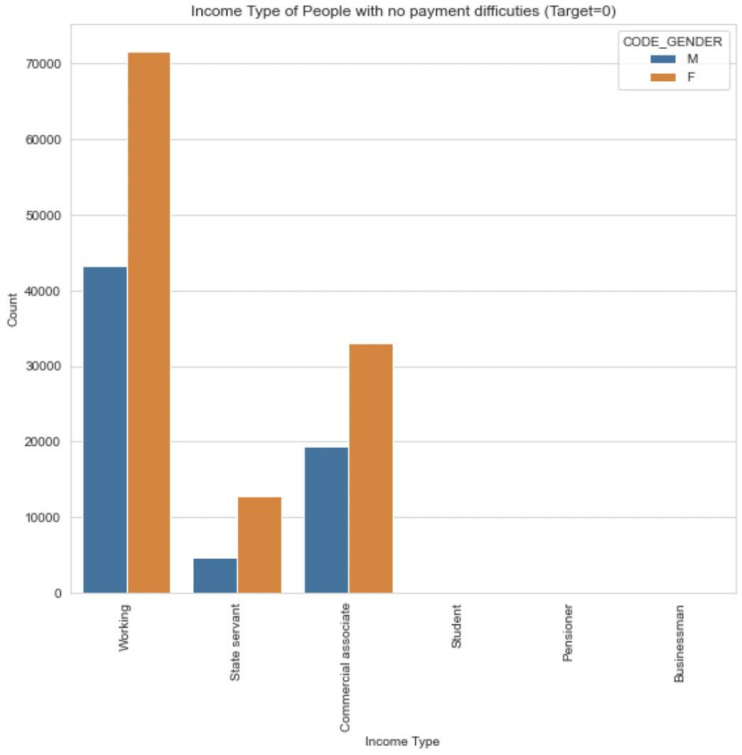
- `125000-150000` is the range with the highest count of values in both `target0` and `target1`





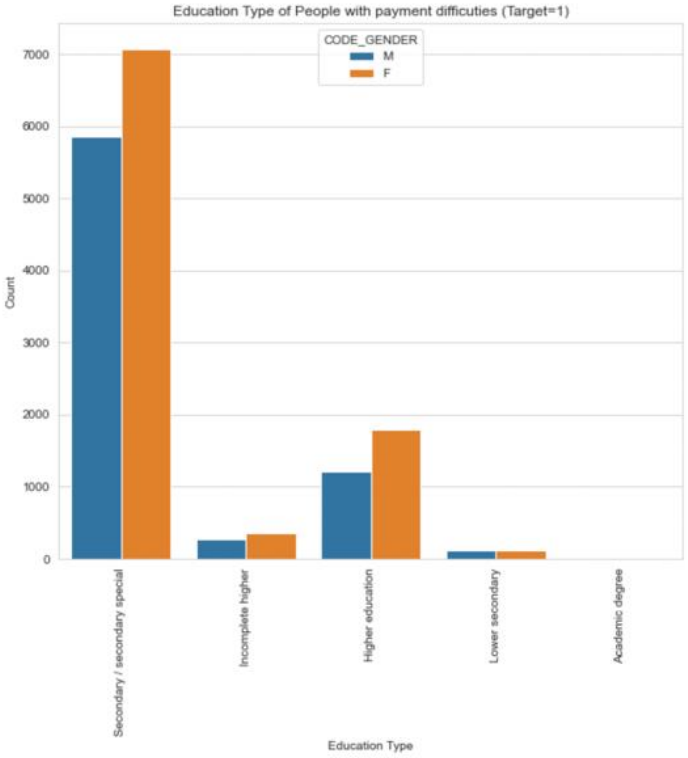
- If we see the Loan Amount Credit-Range `target0` have higher number of amount credit from `200000-550000` and `target1` have lesser number of amount credit in the same range, it means both the targets have maximum number of requirement of loan in this range.
- Loan Amount Credit-Range of `900000 and above` people shows highest number of requirement in their respective targets.

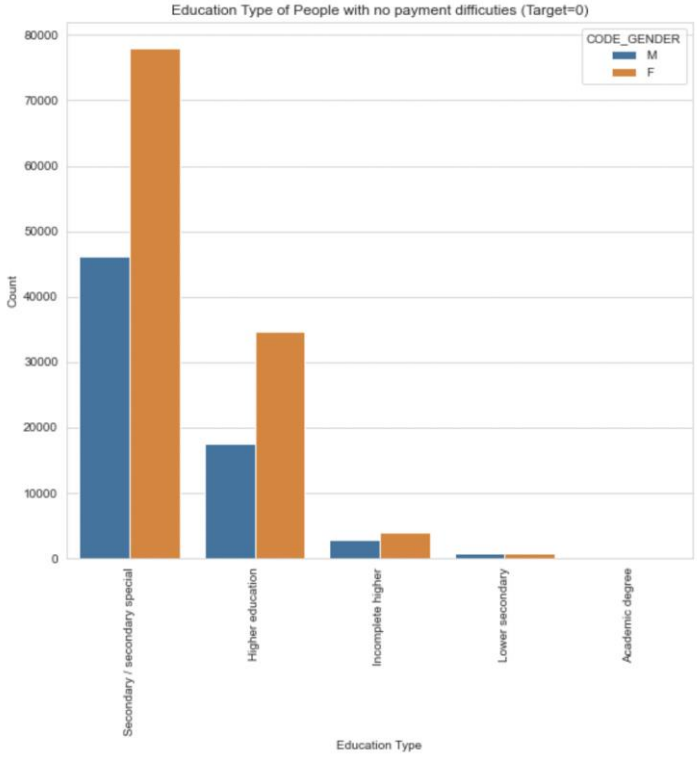




- The `Student`, `Pensioner` and `Businessman` are present in the income type of `target0` but they are absent in `target1`.

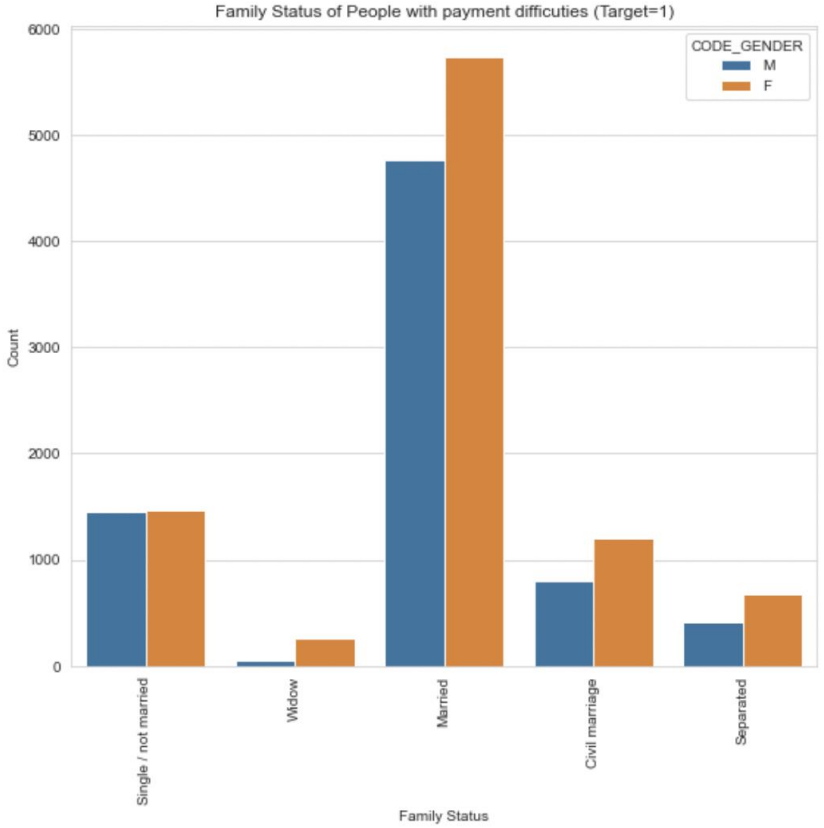
- The number of Females is high as compared to Males in the income type of `Working, State Servant, and Commercial Associate` irrespective of the targets.

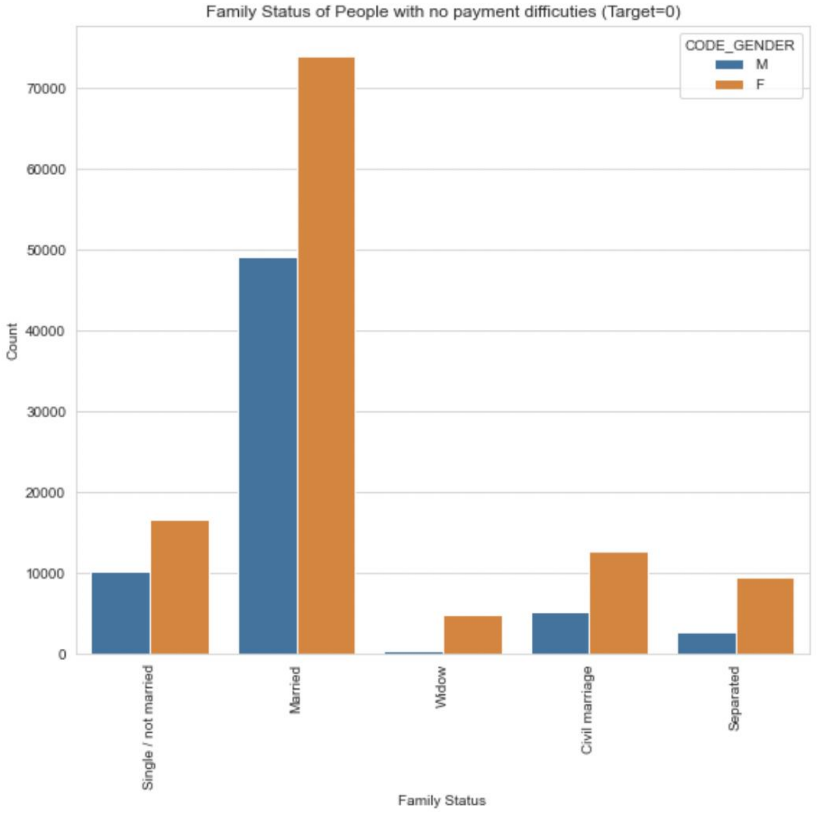

Education Type of People with payment difficulties (Target=1)
CODE_GENDER
M
F
Count
0
1000
2000
3000
4000
5000
6000
7000
Secondary / secondary special
Incomplete higher
Higher education
Lower secondary
Academic degree
Education Type


Education Type of People with no payment difficulties (Target=0)
CODE_GENDER
M
F
Count
0
10000
20000
30000
40000
50000
60000
70000
80000
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
Education Type

- In both the `target0` and `target1` we have the highest number of application from Secondary Special and lowest number of application from Academic Degrees.
- Distribution of `Females` in both the `target0` and `target1` is high in all types of Education.

- In `target0` we can see the distribution of Males are decreased as compared to Females in all types of Education.
- In `target1` we can see the distribution of Males and Females in all types of Education have not much difference in application.


Family Status of People with payment difficuties (Target=1)
CODE_GENDER
M
F
Count
6000
5000
4000
3000
2000
1000
0
Single / not married
Widow
Married
Civil marriage
Separated
Family Status


Family Status of People with no payment difficuties (Target=0)
CODE_GENDER
M
F
Count
70000
60000
50000
40000
30000
20000
10000
0
Single / not married
Married
Widow
Civil marriage
Separated
Family Status

- In both the `target0` and `target1` the highest distribution are of Married peoples and the lowest distribution is of Widow's, irrespective of their gender.

- We have Females in `target1` in higher numbers as compared to Males, which means Females of all type of Family Status are more likely to default.

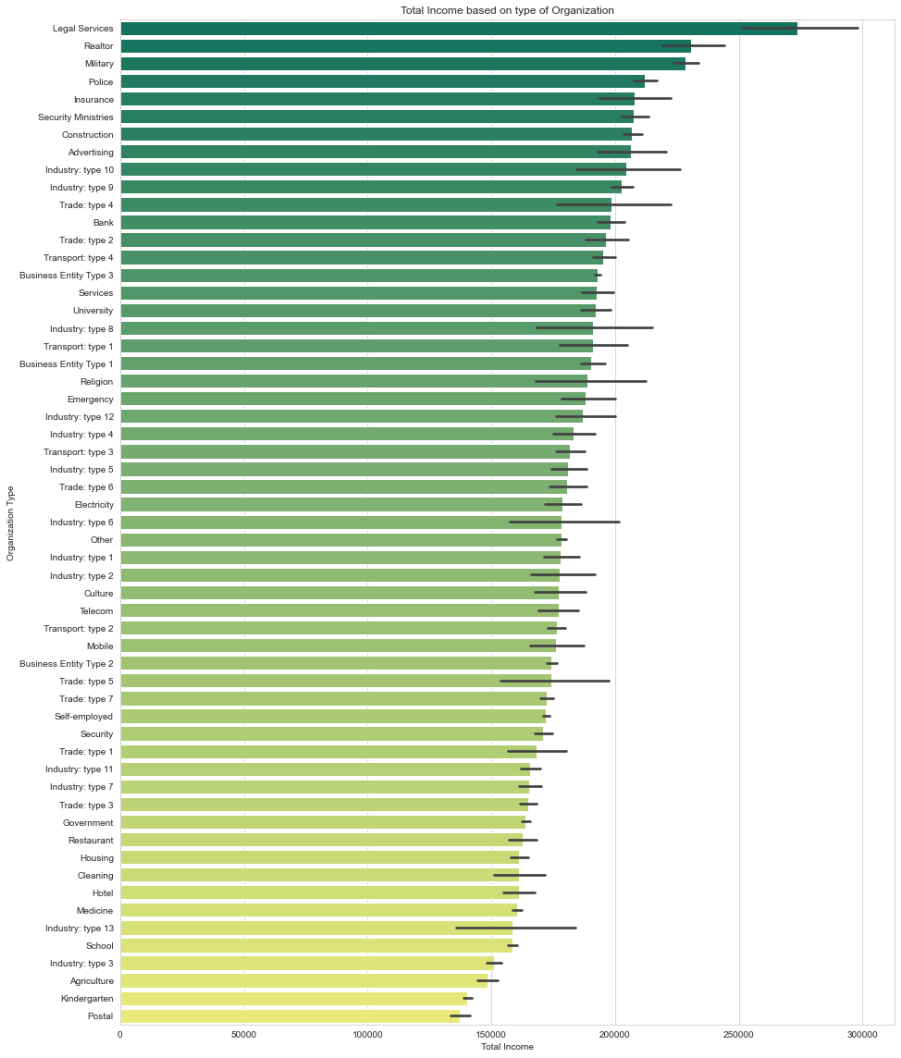
- We can also see in `target0` the the distribution of Females are higher as compared to Males, so that's why the `target1` have also high distribution of Females.

Bivariate Analysis on New Application

- The highest Total Income in Organization Type is of `Legal Services` at around 275000 in `target0`.

- The lowest Total Income in Organization Type is of `Postal` which is less than 150000.

- Although we have the highest number of application from `Business Entity Type-3` but it's Total Income is lower than the `Legal Services`.

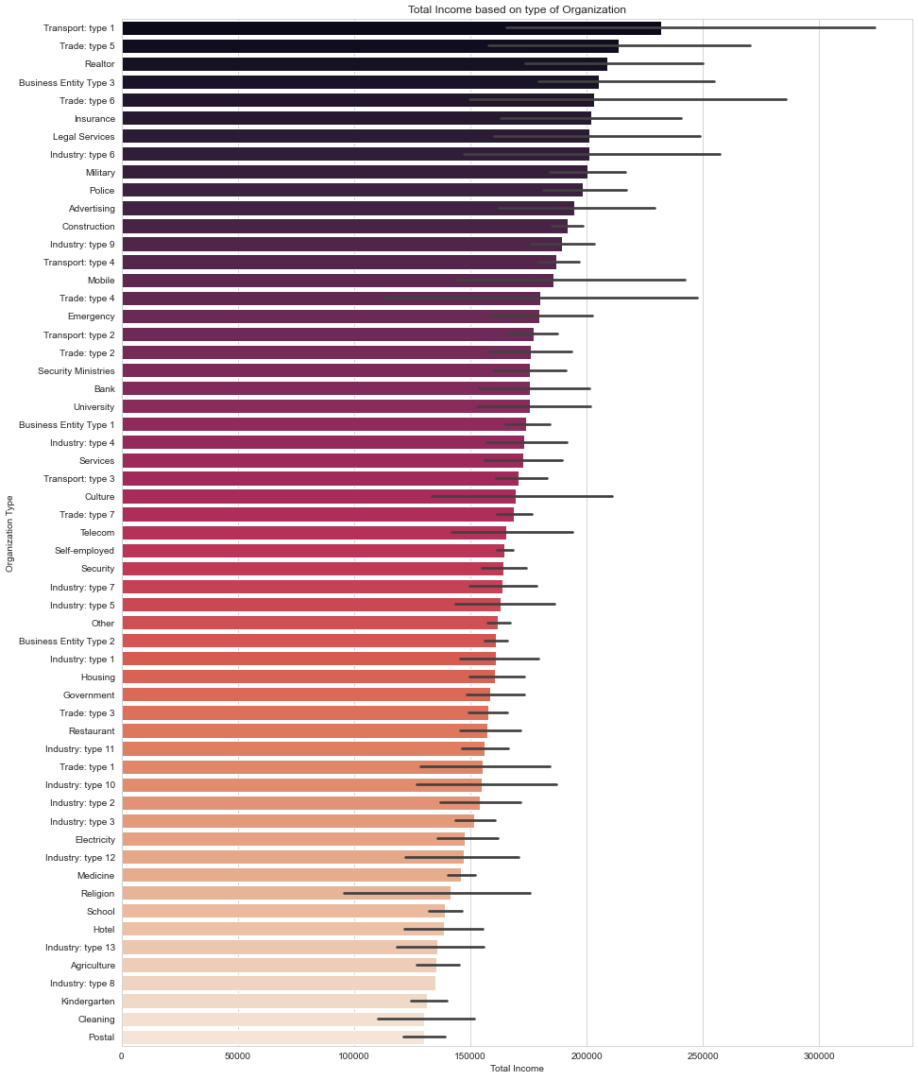


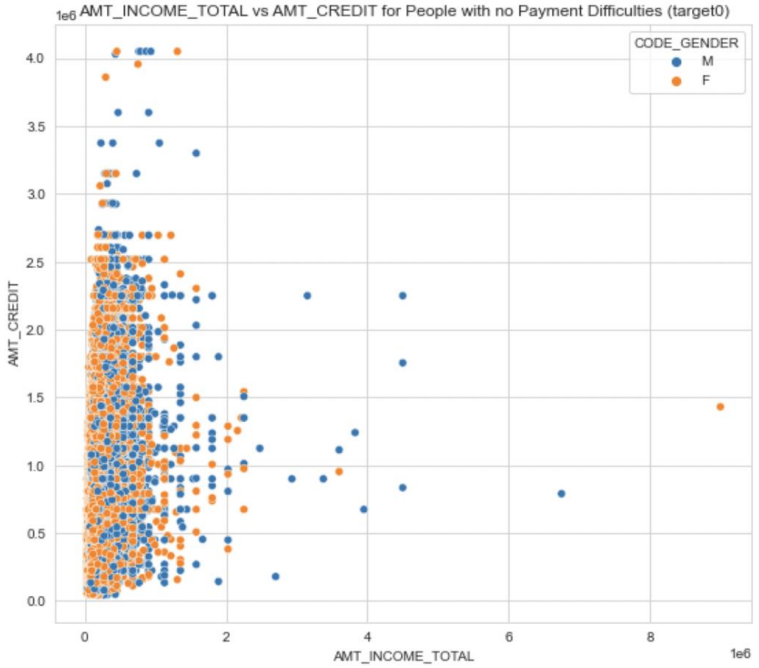
- The highest Total Income in Organization Type is of `Transport Type 1` at about 250000 in `target1`.

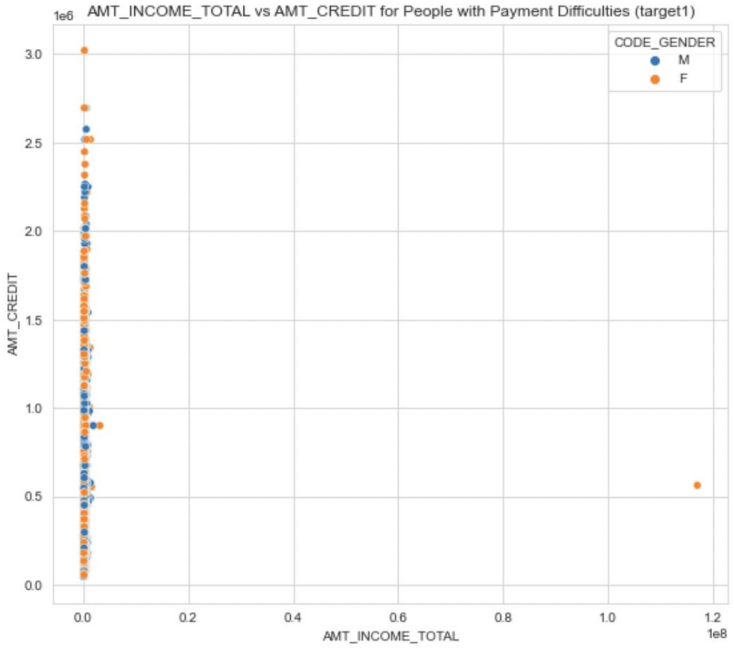
- The lowest Total Income in Organization Type is of `Postal and Cleaning` which is less than 150000.

- Although we have the highest number of application from `Business Entity Type-3` but it's Total Income is lower than the `Transport Type 1`.

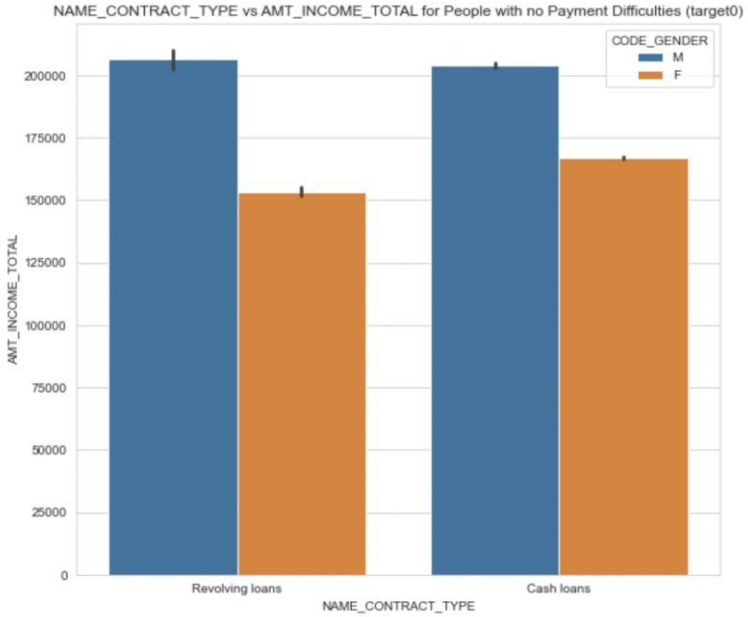
- In target1, `Transport Type 1` has the highest total income amount, whereas in target0, `Legal Services` are leading in terms of total income. This can mean that people with payment difficulties might be majorly from `Transport Type 1` and their average total income will be higher as compared to other types of organizatons.

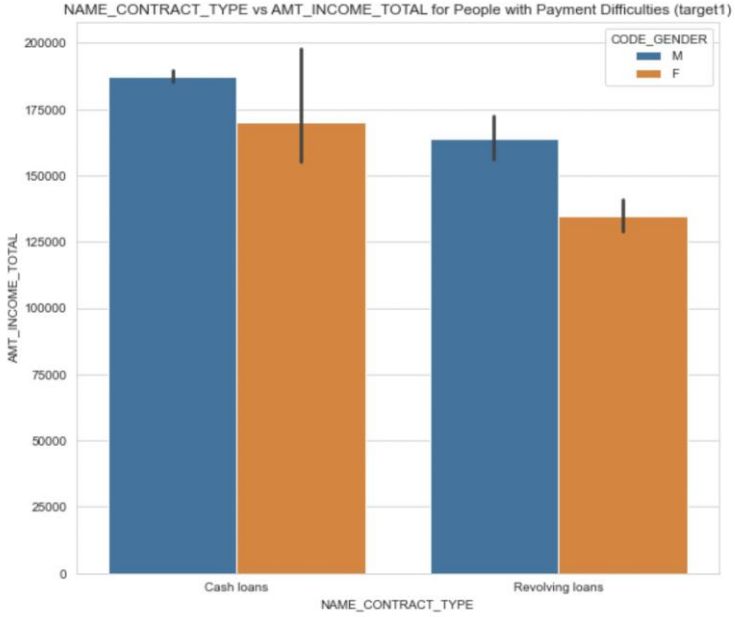




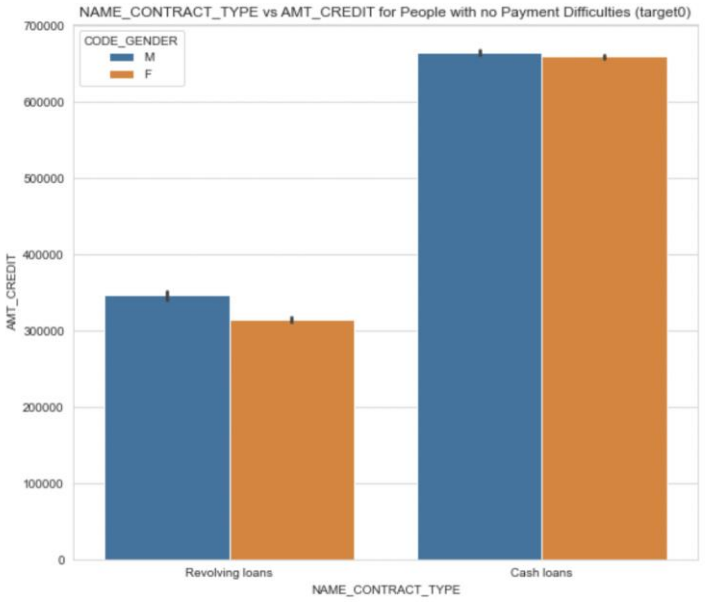


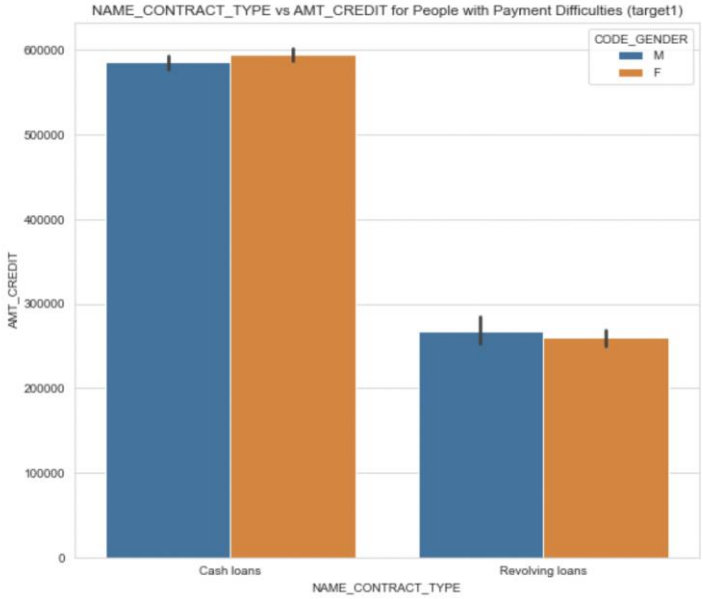
- In `target0`and `target1` higher Amount of Income does not mean higher Amount of Loan Credit.
- The `target0` have both Males and Females losely together.
- The `target1` have both Males and Females in a linear correlation with only few outliers.
- Most of the high Loan Amount Credit are provided to the low Amount of Income Total applicants.



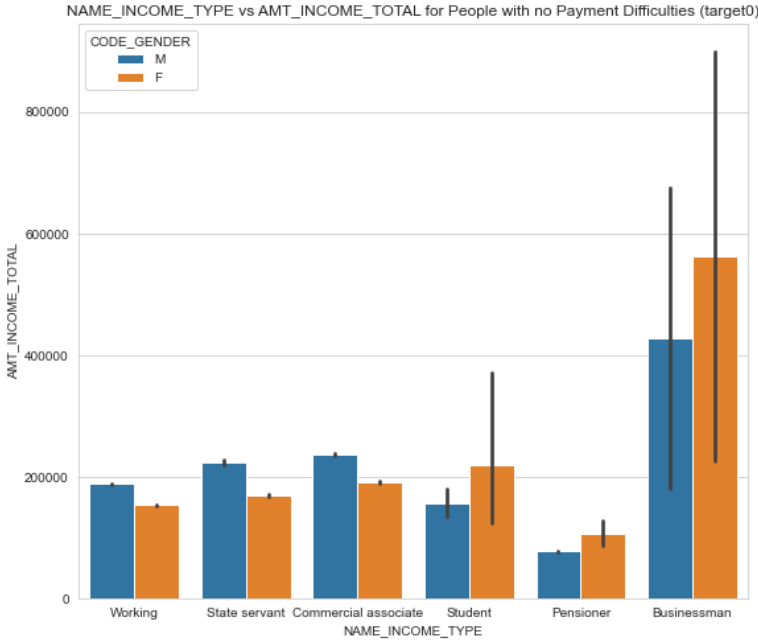


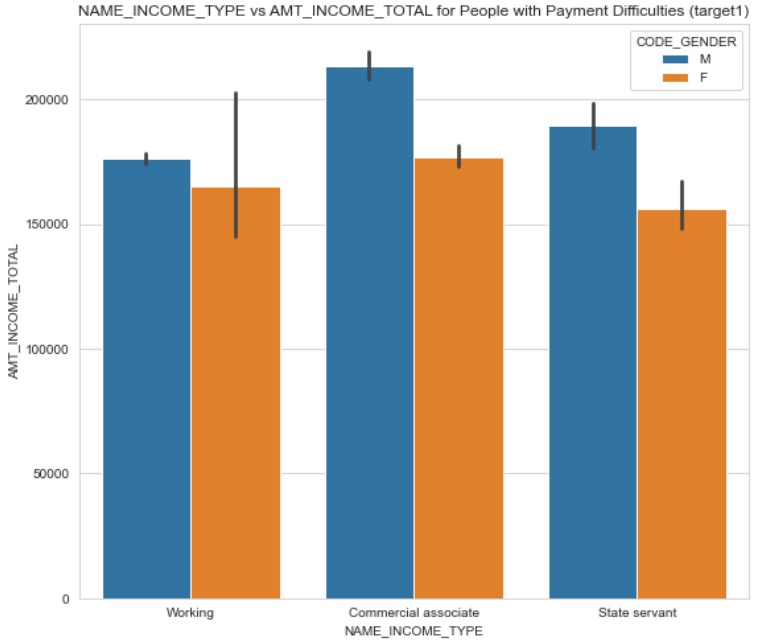
- In `target0` the Income Total of applicants is quite similar in both the `Revolving Loans and Cash Loan`, by irrespective of their gender.
- Where as in `target1` we can infer that we have lesser Income Total of applicants in `Revolving Loans` as compared to `Cash Loans, by irrespective of their gender.
- The difference between the count of values in `Revolving Loans` and `Cash Loans` is significantly higher in both the `target0` and `target1`, but their data about the total income amount is almost equal with slight differences.



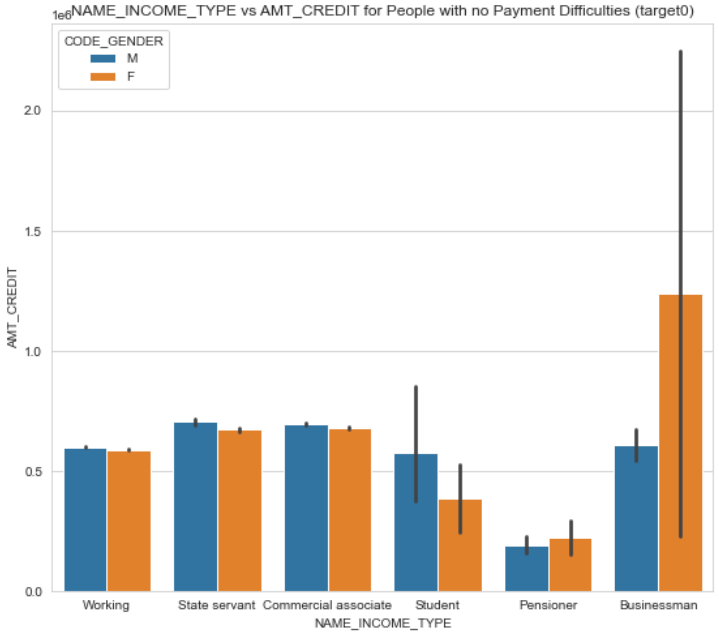


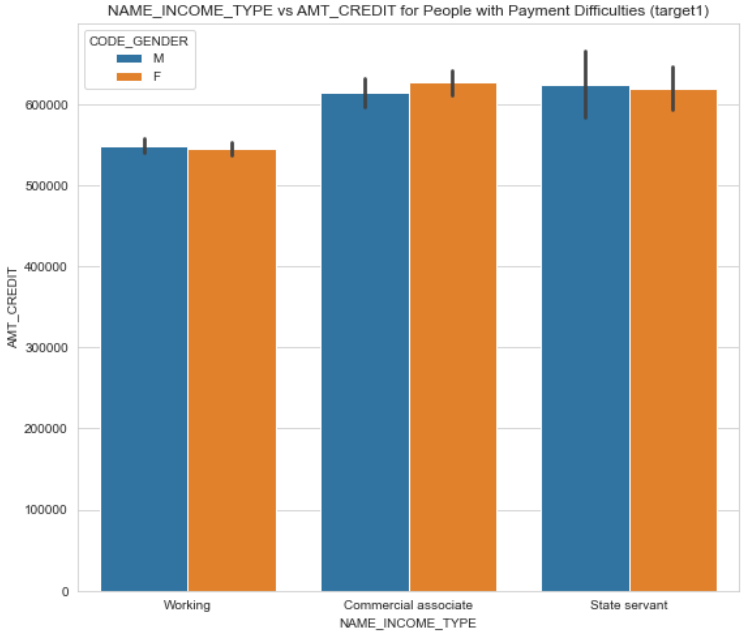
- In `target0` and `target1` the Loan Credit Amount is much higher in `Cash Loans` as compared to `Revolving Loans`, and if we compare gender of both the target, then in `target0` Males have been provided slightly more with the Cash Loan as compared to Females, whereas in `target1` it is vice-versa.



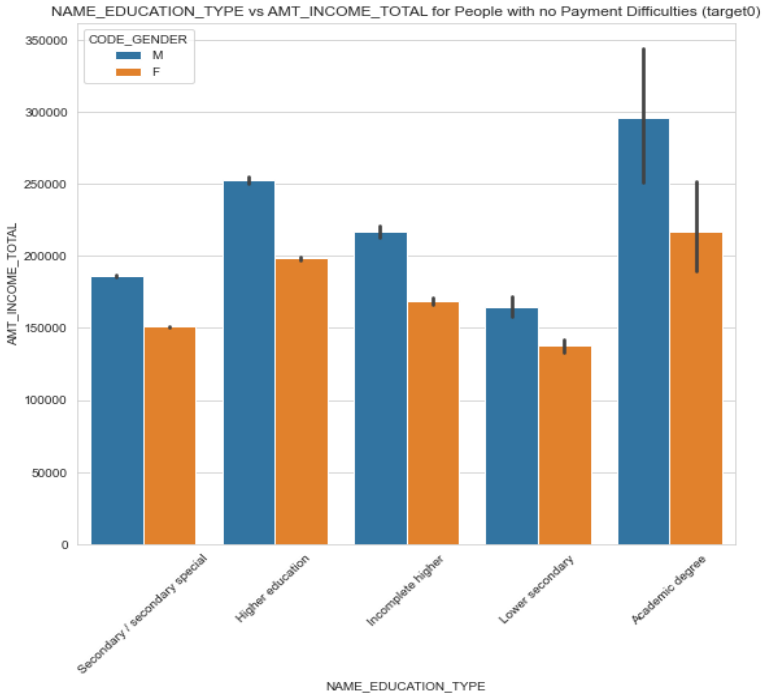


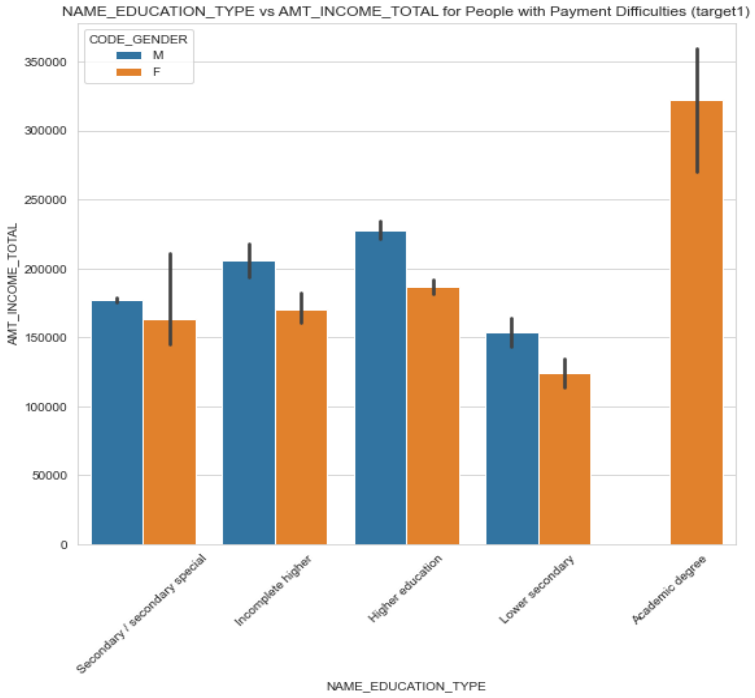
- In `target0` the Businessman Income Type have highest Amount Income, whereas this Income type is absent in `target1`.
- In `target0` the Pensioner Income Type have the lowest Amount Income, whereas this is also absent in `target1`.
- In `target1` the Commercial Associate have the highest Amount Income and Working Income Type is the lowest in Amount Income.



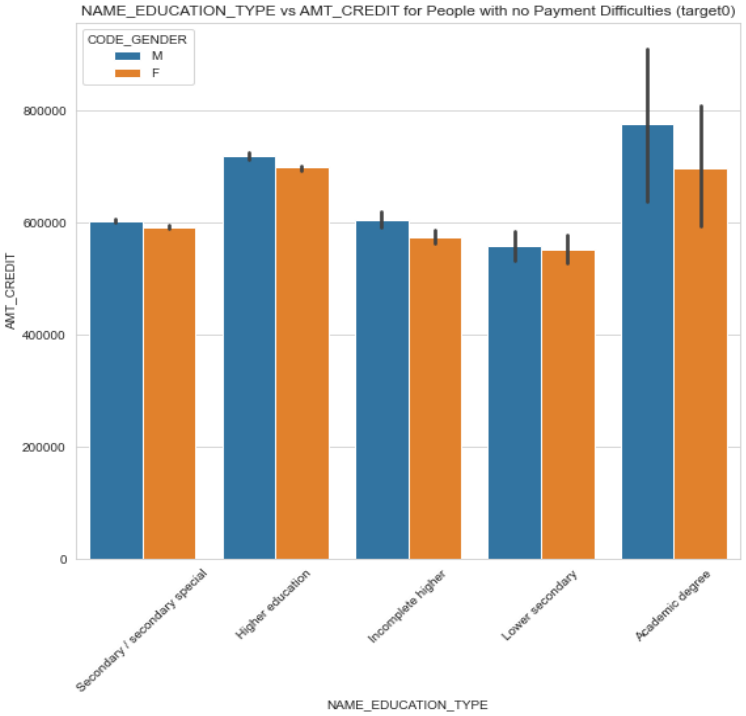


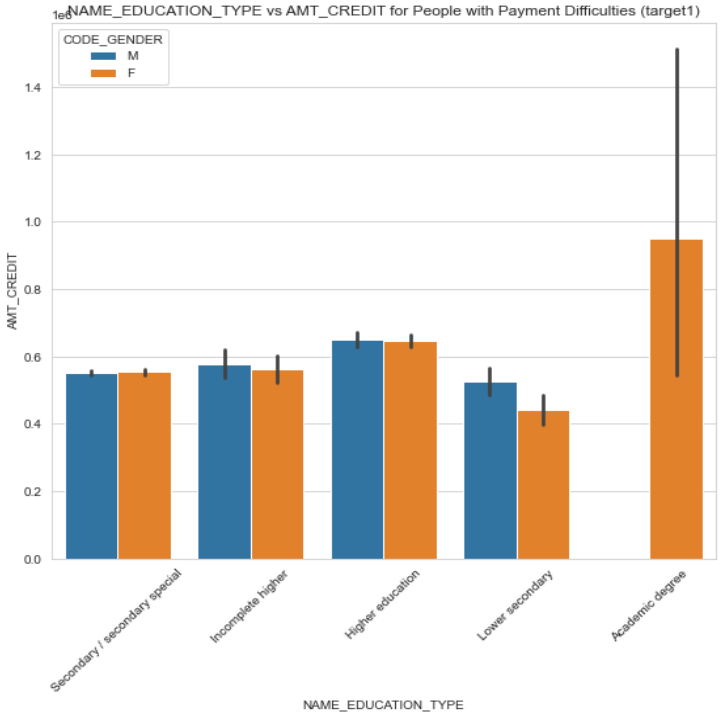
- In `target1` the Commercial Associate and State Servant are slightly equal in Loan Amount Credit which is greater than `600000`.
- In `target0` the Businessman Income Type have the highest Loan Amount Credit in Females, whereas in Males the State servant have highest Loan Amount Credit.
- In `target0` the Pensioners are the lowest in Credit Amount.



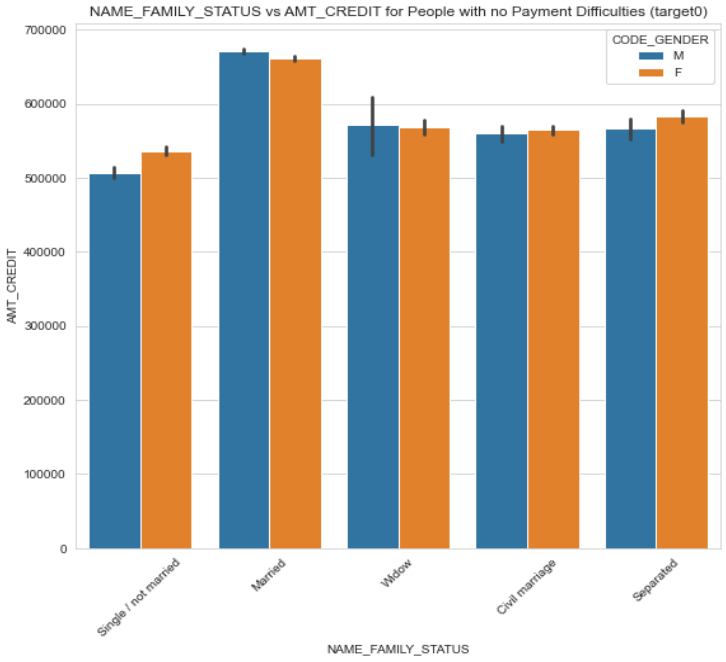


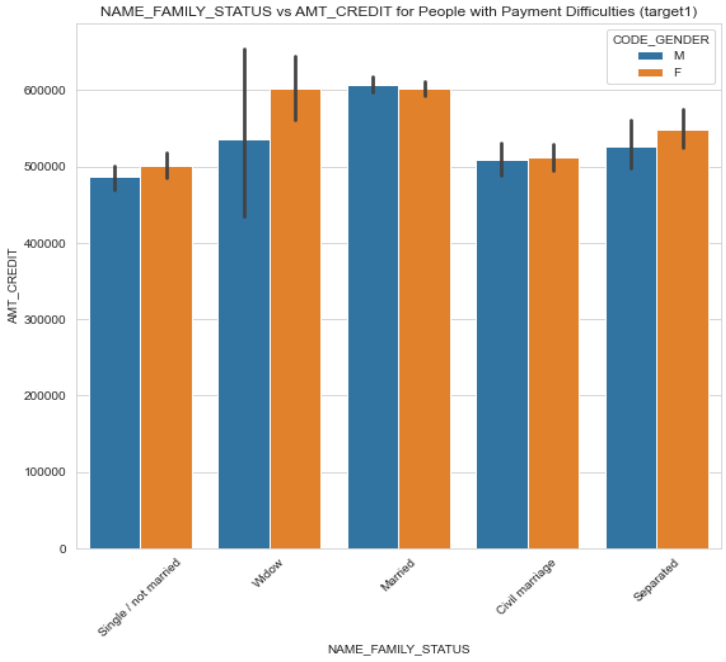
- In `target0`, the Academic Degree have the highest Amount of Income, by irrespective of their gender.
- In `target0`, the Lower Secondary have the lowest Amount of Income, by irrespective of their gender.
- In `target1`, the Higher Education have the highest Amount of Income in Males, whereas Females have highest Amount of Income in Academic Degree.
- Female of Academic Degree from `target1` have more Amount of Income than the `target0` highest Amount of Income.



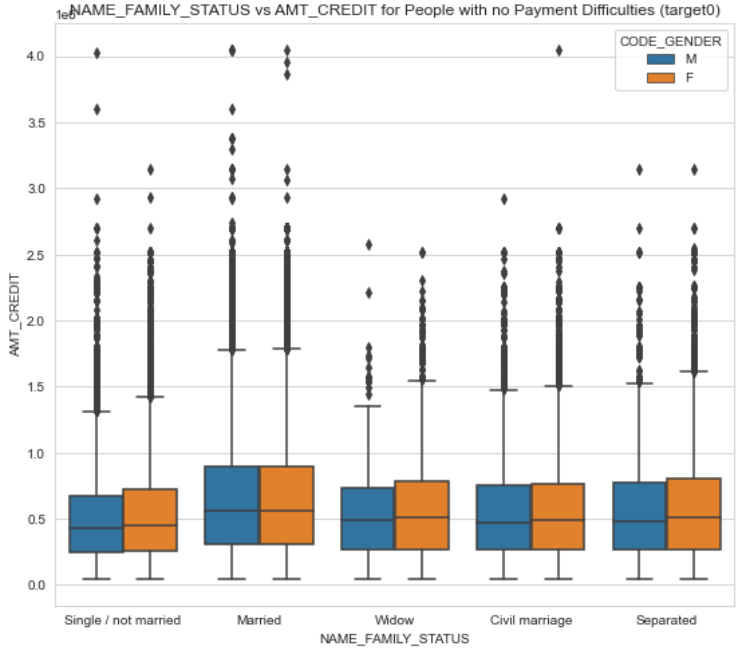


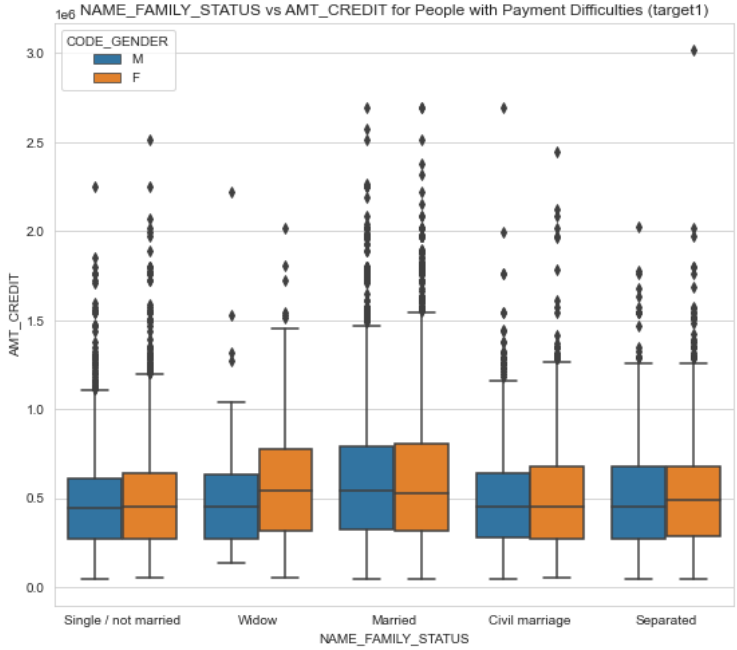
- In both the `target0` and `target1` the Academic Degree and Lower Secondary have the highest and lowest Loan Credit Amount, respectively.
- Males of Academic Degree from `target0` and the Females of Academic Degree from `target1` have slightly similar Loan Credit Amount.





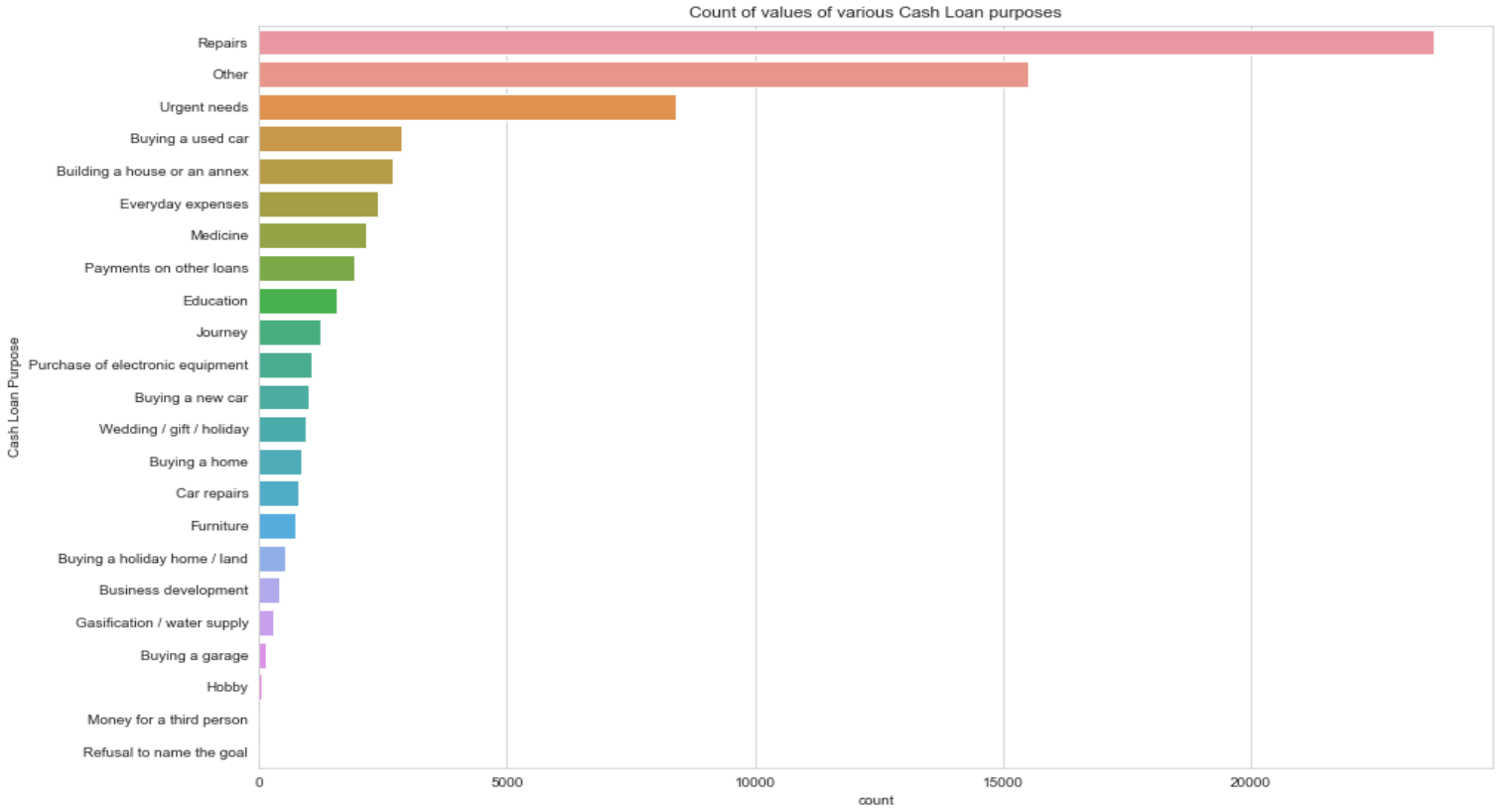
- In `target0` the Married Males and Females have highest Loan Amount Credit, where as Single/Not Married applicants have lowest Loan Amount Credit.
- In `target1` the Married Males and Females have highest Loan Amount Credit, where as Single/Not Married applicants have lowest Loan Amount Credit.
- In `target1`, the Widow Females have more Loan Amount Credit than the `target0` Widow.
- All the other type of Family Status have slightly near to each other in Loan Amount Credit.



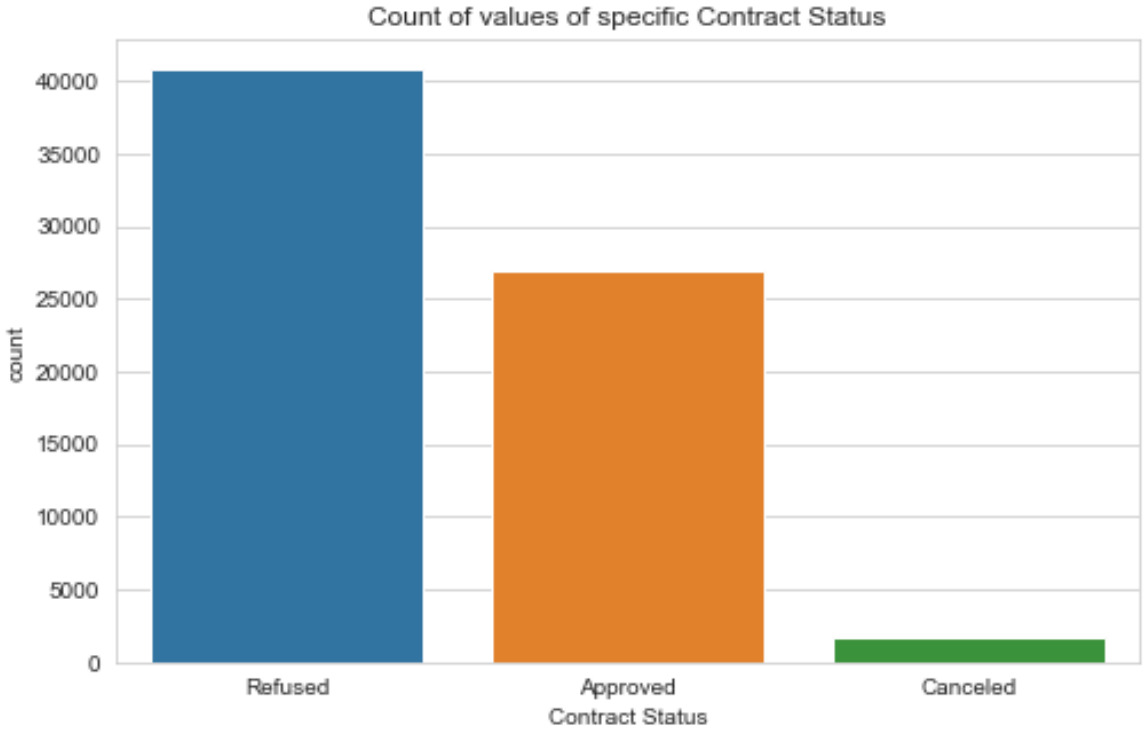


- We can infer that there are outliers in both the `target0` and `target1`, by irrespective of their gender.
- In `target0` Married, Single, Civil Marriage have slightly bigger third quartile as compared to other quartiles, irrespective of their gender.
- In `target1` Married, Single, Civil Marriage have slightly bigger third quartile as compared to other quartiles, irrespective of their gender.
- In `target0` the third quartile of `Separated` Females is bigger than its first quartile, whereas it is vice-versa in `target1`.

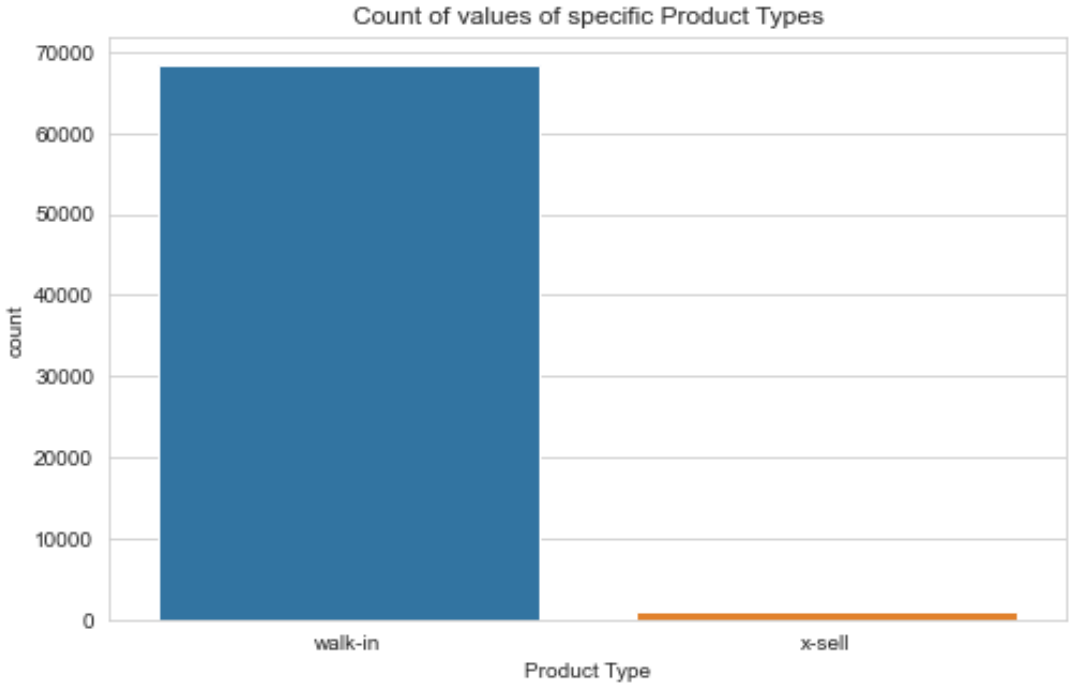
Univariate Analysis on Previous Application



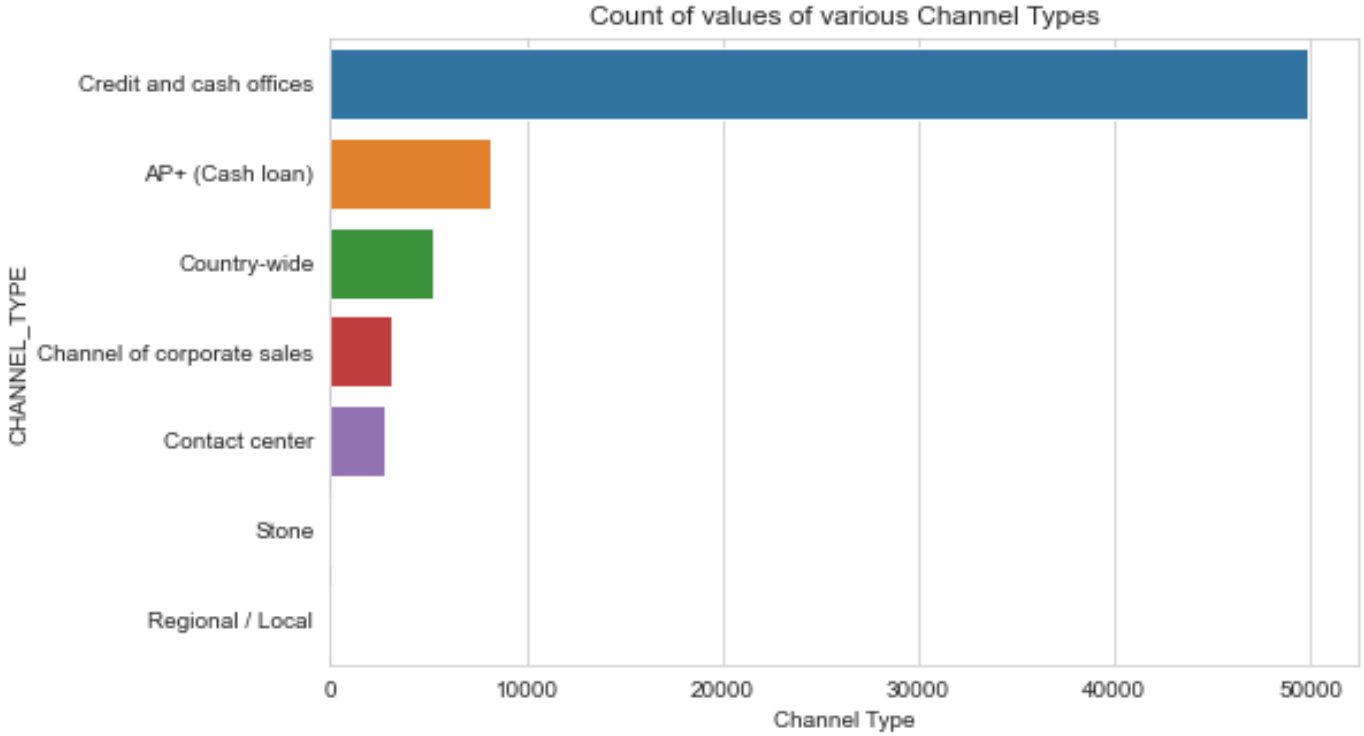
- The highest applications are for the repair purposes.
- The most applications are for the purpose of `repairs`, `other` and `Urgent Needs`.
- It is surprising to see that count of people who applied buying a home are low.



- Most of the applications are being refused.
- There is a considerable gap between approved and refused applications as the number of approved applications are significantly lower than the refused applications count.
- Very Low count of applications are being cancelled.

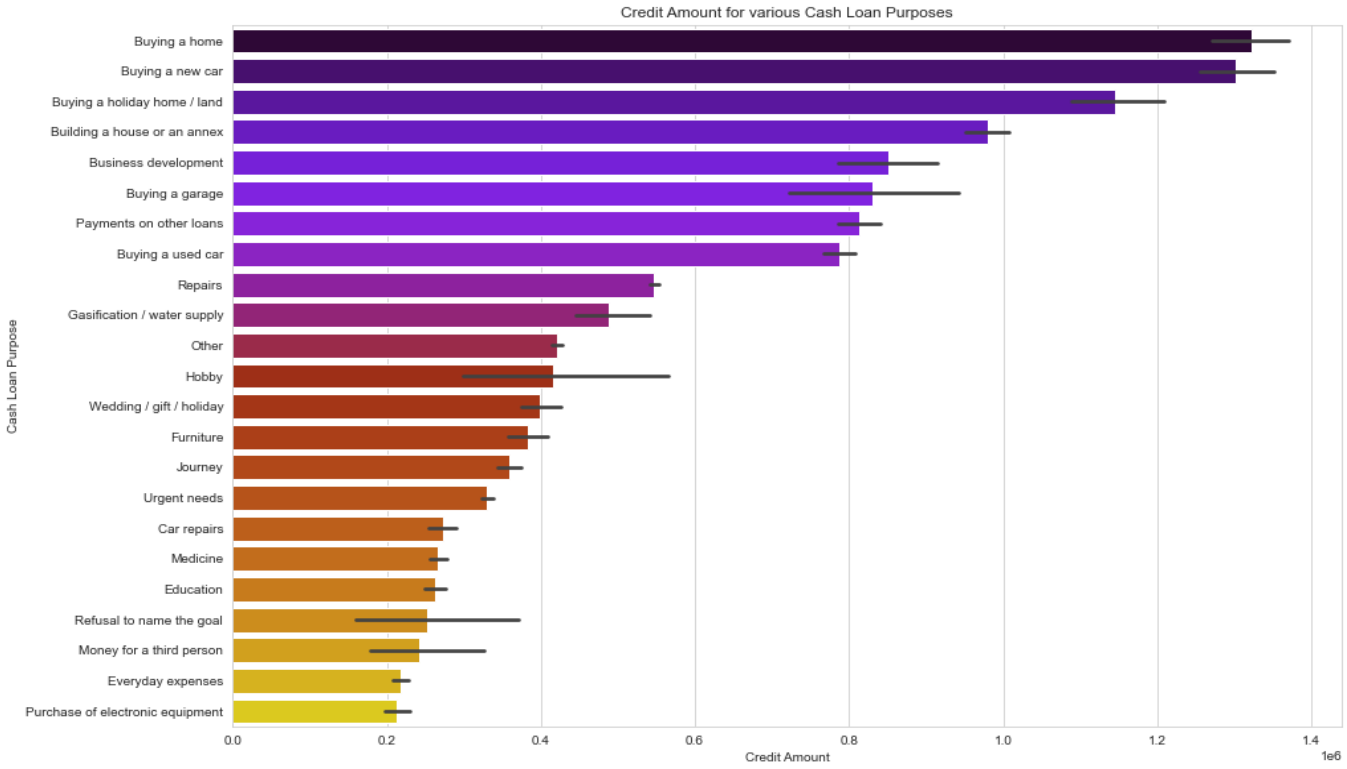


- Almost all the people who applied for the loan were `walk-in` customers
- `X-sell` applications were very few.

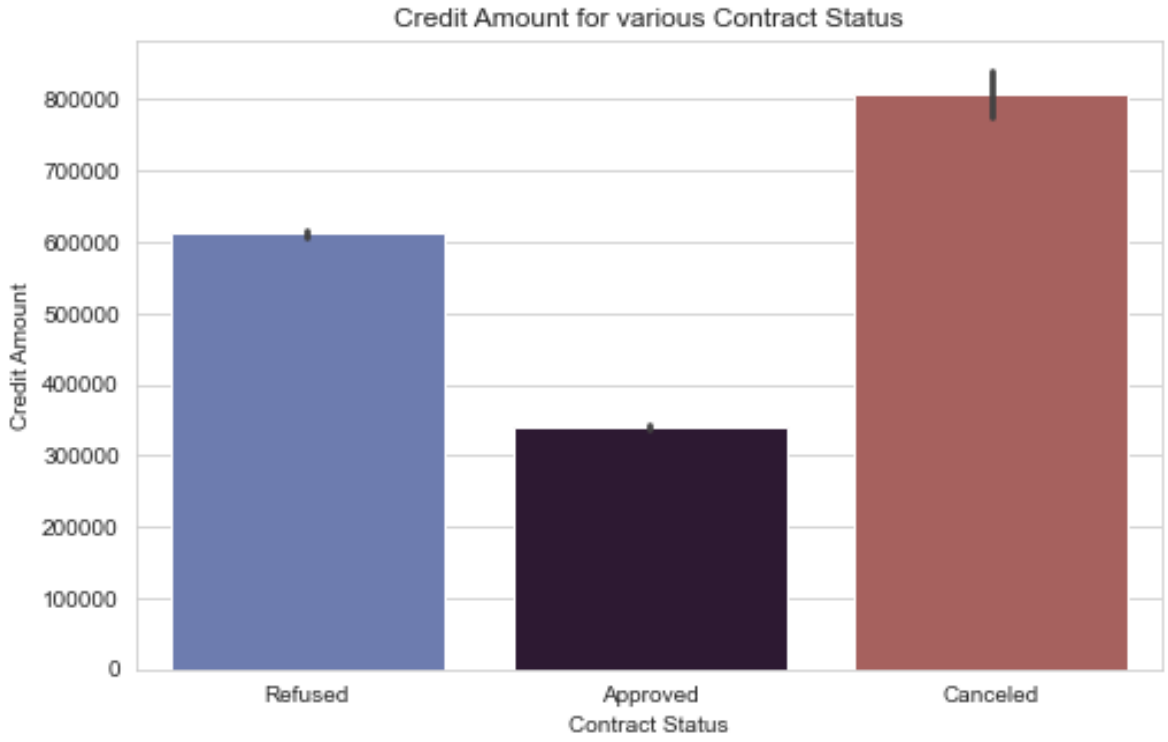


- `Credit and Cash offices` is the channel type with the highest count of values
- `Stone` and `Regional/Local` channels have very low count of values.

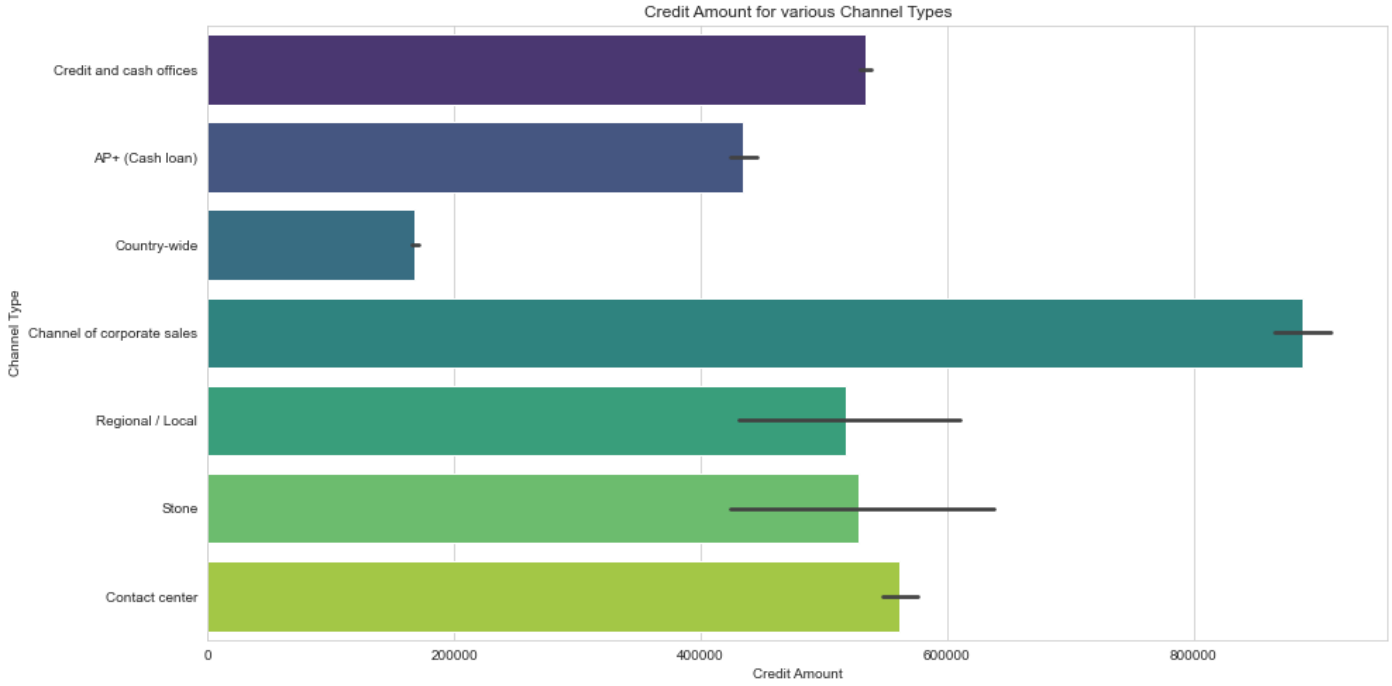
Bivariate Analysis on Previous Application



- The average Credit Amount for buying a new car, building a house or an annex, other loan payments, buying used car, buying a holiday home/land, buying a garage, and developing business is on the higher side.
- The lowest average Credit Amount is for the purchase of electronic equipment.

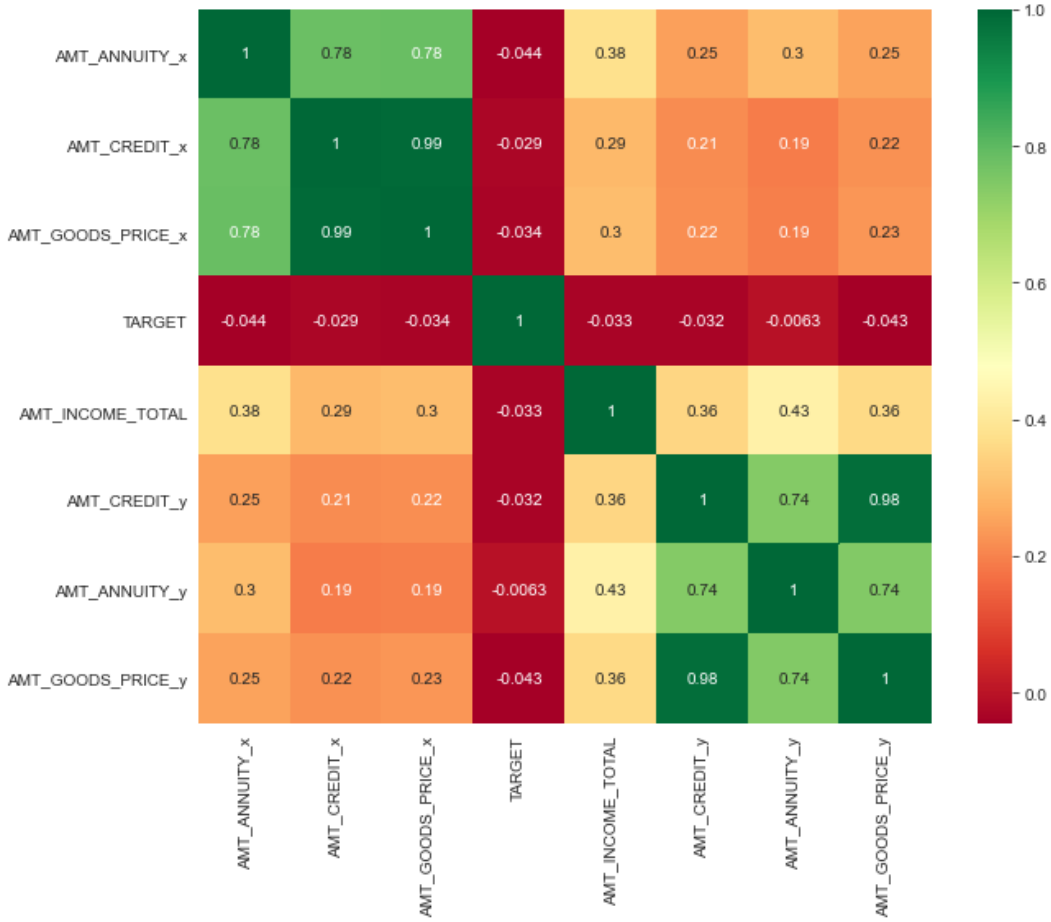


- The Credit Amount for `Cancelled` Contracts are also the highest followed by the `Refused` contracts.
- The `Approved` contracts have the lowest Annuity Amount here as well

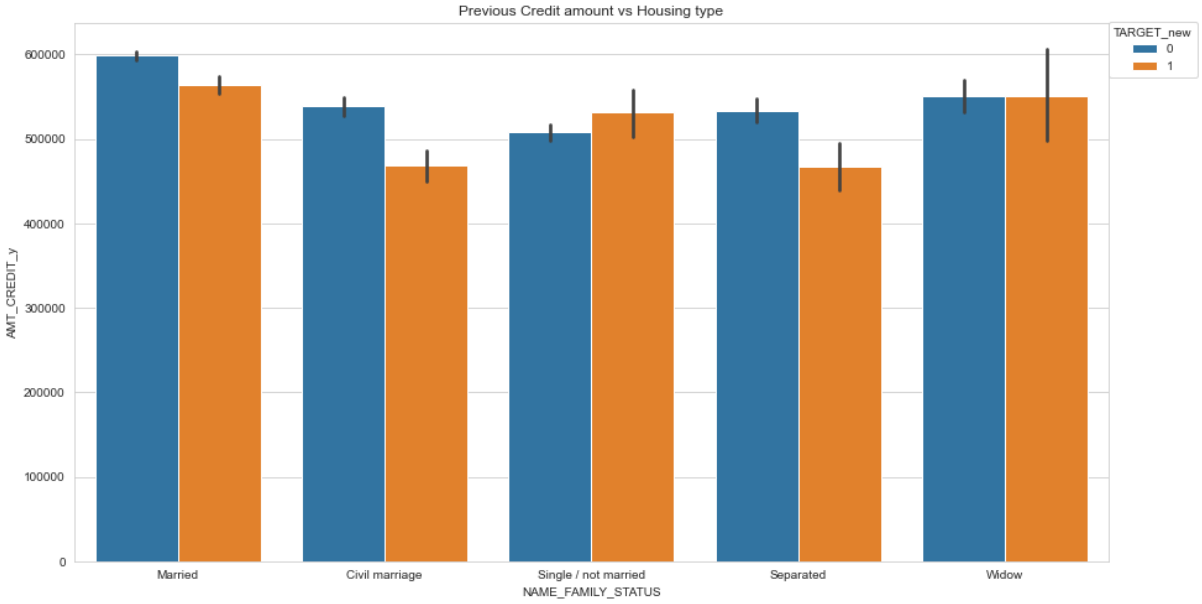


- The Credit Amount for `Channel of the corporate sales` is the highest.
- The `Country-wide` have the lowest Credit Amount as well.
- Although the count of values for `Credit and Cash Offices` is higher, but the credit amount is higher in `Channel of Corporate Sales`.

Bivariate Analysis on Merged Dataset

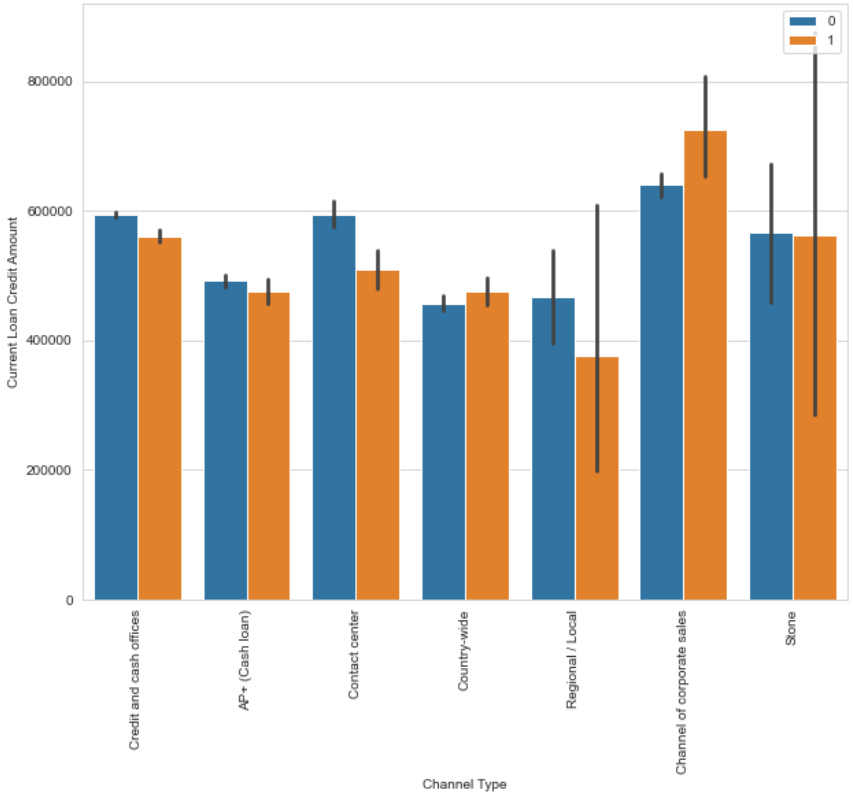

AMT_ANNUITY_x
AMT_CREDIT_x
AMT_GOODS_PRICE_x
TARGET
AMT_INCOME_TOTAL
AMT_CREDIT_y
AMT_ANNUITY_y
AMT_GOODS_PRICE_y
1 0.78 0.78 -0.044 0.38 0.25 0.3 0.25
0.78 1 0.99 -0.029 0.29 0.21 0.19 0.22
0.78 0.99 1 -0.034 0.3 0.22 0.19 0.23
-0.044 -0.029 -0.034 1 -0.033 -0.032 -0.0063 -0.043
0.38 0.29 0.3 -0.033 1 0.36 0.43 0.36
0.25 0.21 0.22 -0.032 0.36 1 0.74 0.98
0.3 0.19 0.19 -0.0063 0.43 0.74 1 0.74
0.25 0.22 0.23 -0.043 0.36 0.98 0.74 1
1.0
0.8
0.6
0.4
0.2
0.0

- `AMT_GOODS_PRICE_y` and `AMT_CREDIT_y` have positive linear correlation
- `AMT_GOODS_PRICE_y` and `AMT_ANNUITY_y` also have positive linear correlation
- `AMT_GOODS_PRICE_y` and `TARGET` have negative correlation
- `AMT_ANNUITY_y` and `AMT_CREDIT_y` have positive correlation
- `AMT_GOODS_PRICE_x` has strong positive linear correlation with `AMT_CREDIT_x` and `AMT_ANNUITY_x`
- `AMT_ANNUITY_x` and `AMT_CREDIT_x` also have positive correlation
- All of the variables plotted above have low correlation with `TARGET`

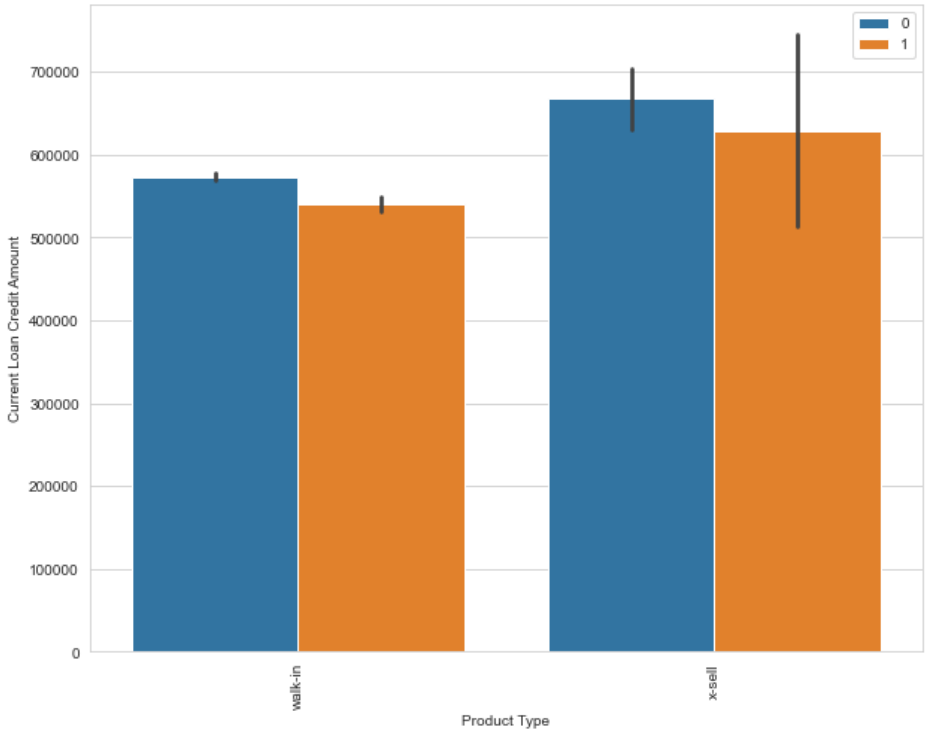


- `Married` people have the highest average credit amount in both the `target0` and `target1`
- The average credit amount for widows in both `target0` and `target1` is equal
- The plots for `Civil marriage` and `Separated` people is almost similar for in both `target0` and `target1`

- `Channel of Corporate Sales` have highest number of people with payment difficulties according to their current Credit Amount.
- - `Regional/Local` have the lowest number of people with payment difficulties according to their current Credit Amount.



- `walk-in` customers have the lowest average credit amount for both `target0` and `target1`
- - `x-sell` customers have the highest average credit amount for both `target0` and `target1`



CONCLUSION

1. There is a high chance of `Females` becoming defaulters as compare to `Males`. So, we can increase the `Cash Loan` credit to `Males` because they are least likely to be present in `target1`.

2. Our `Business Entity: Type-3` is the most highest type of Organization which have defaulters, instead of this we can focus on `Bank`, `Advertising`, `Hotel`, to expand the business of the bank, because they are less likely to default.

3. We can also consider `Legal Services` for banks business, because they have the highest Amount of Income and less chances of default as compare to `Business Entity: Type-3`.

4. We can focus on `Males` of `Academic Degrees`, because they have a good Amount of Income and they are very less likely to default.

5. We have the least number of applications from `Channel of Corporate Sales`, and these applicants have the highest number of Loan Credit Amount, so for the sake of business of bank, we must increase the application of channel type `Corporate Sales`, as compared to `Credit and Cash Offices`.

6. Also, we have the highest application from `walk-in` customers, but the loan amount credit is higher in `x-sell`, we can increase the application from `x-sell` product type to increase the banks business.

THANK YOU