CREDIT EDA ASSIGNMENT

By : RAGHUNATH V P

PROBLEM STATEMENT

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history.
- Because of that, some consumers use it to their advantage by becoming a defaulter.
- 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 data given contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:
  - **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
  - **All other cases:** All other cases when the payment is paid on time.
- When a client applies for a loan, there are four types of decisions that could be taken by the client/company):
  - **Approved:** The Company has approved loan Application
  - **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.
  - **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
  - **Unused offer:** Loan has been cancelled by the client but at different stages of the process.

This case study will use EDA to understand how consumer attributes and loan attributes influence the tendency to default.

In other words, 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 utilize this knowledge for its portfolio and risk assessment.

DATA SET AVAILABLE:

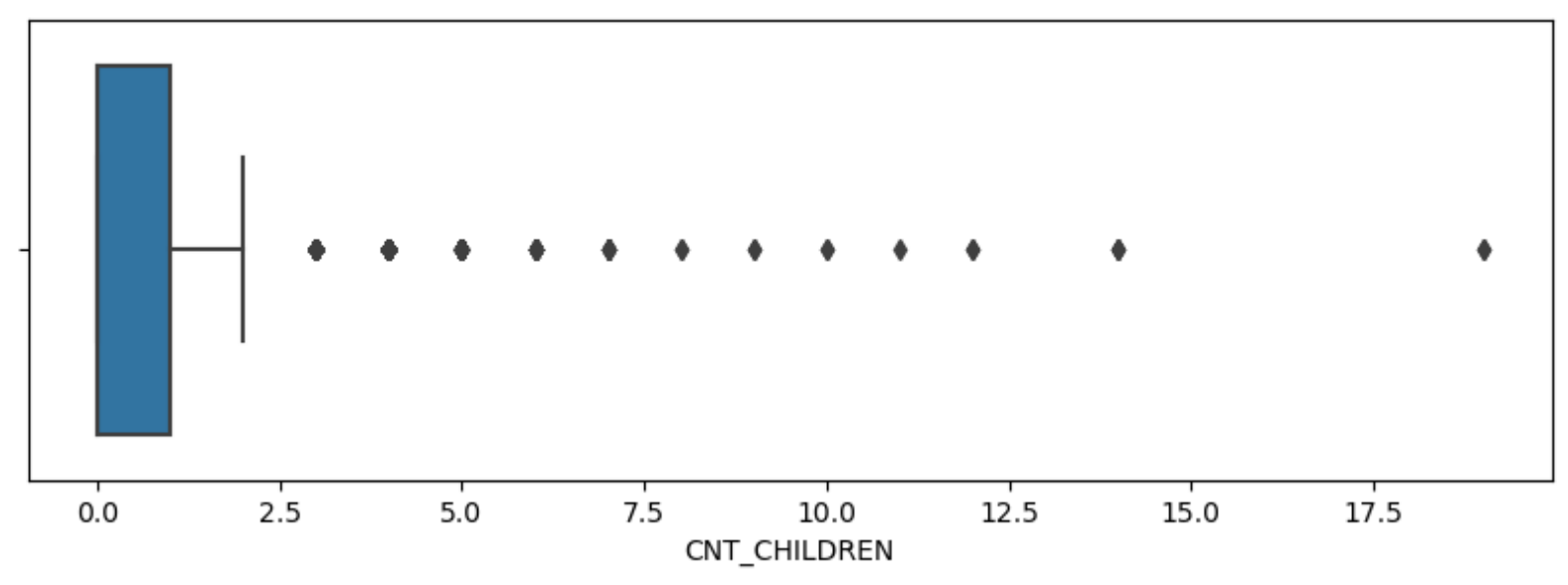
- *1. 'application_data.csv'* contains all the information of the client at the time of application.
  The data is about whether a **client has payment difficulties.**

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

- *3. 'columns_description.csv'* is data dictionary which describes the meaning of the variables.

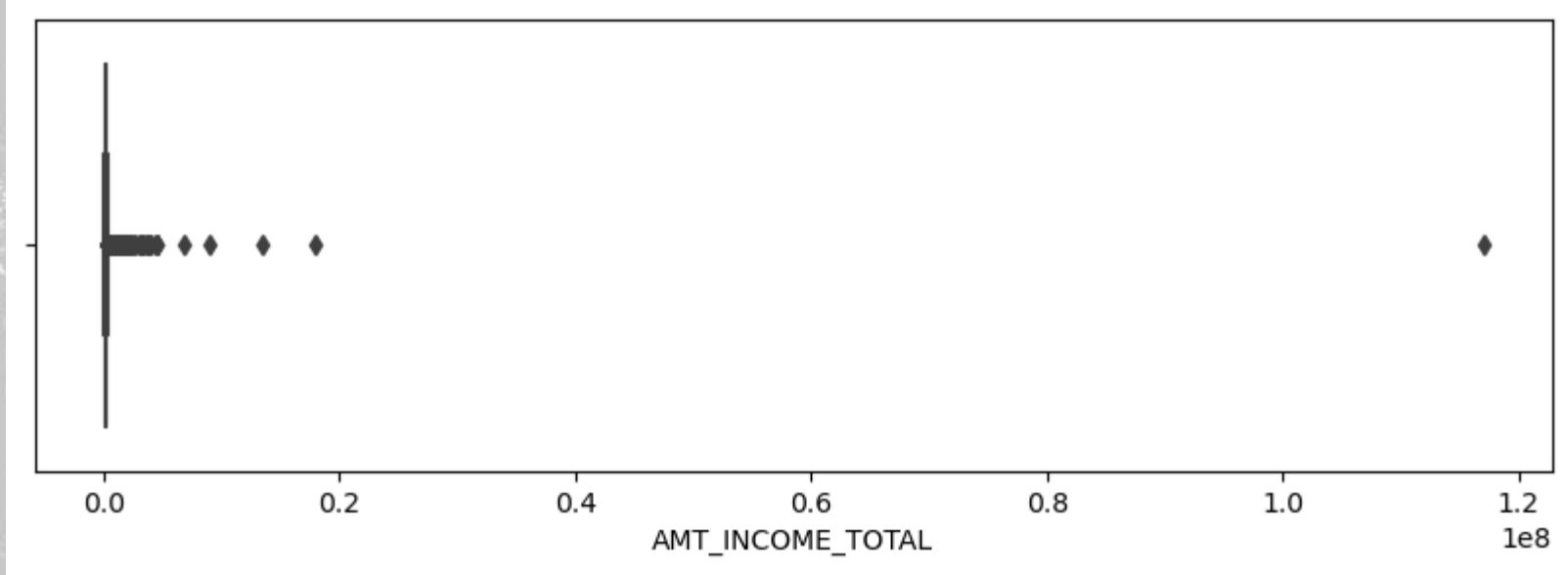
Analysing *'application_data.csv'*

- Checking for outliers in **CNT_CHILDREN**



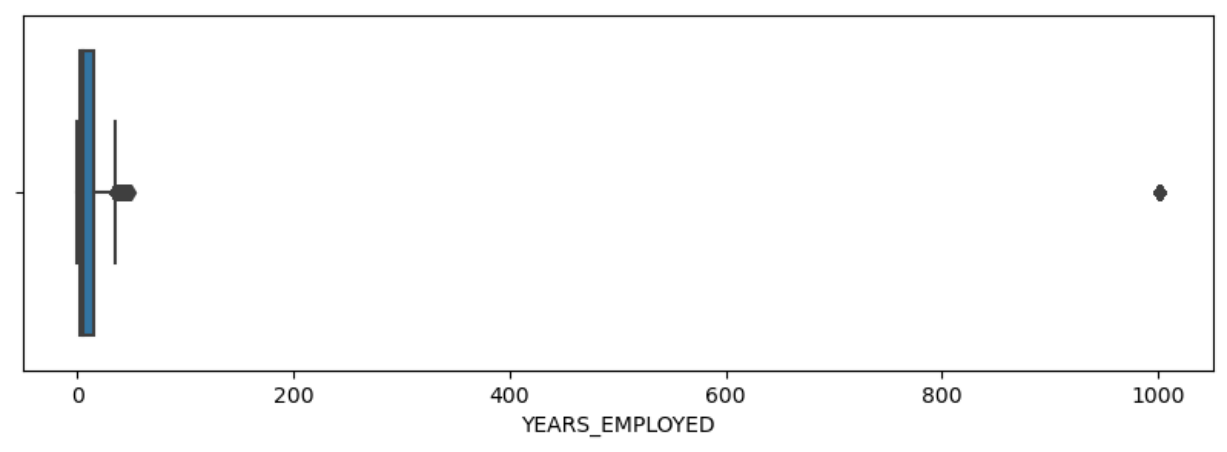
- Most of the values are close to the 1st quartile and there are outliers present.
- CNT_CHILDREN is representing the number of children the client has and there is a range up to 19 for the same which has chance of occurrence.
- So values greater than 3 can be treated as an outlier.

- Checking for outliers in **AMT_INCOME_TOTAL**

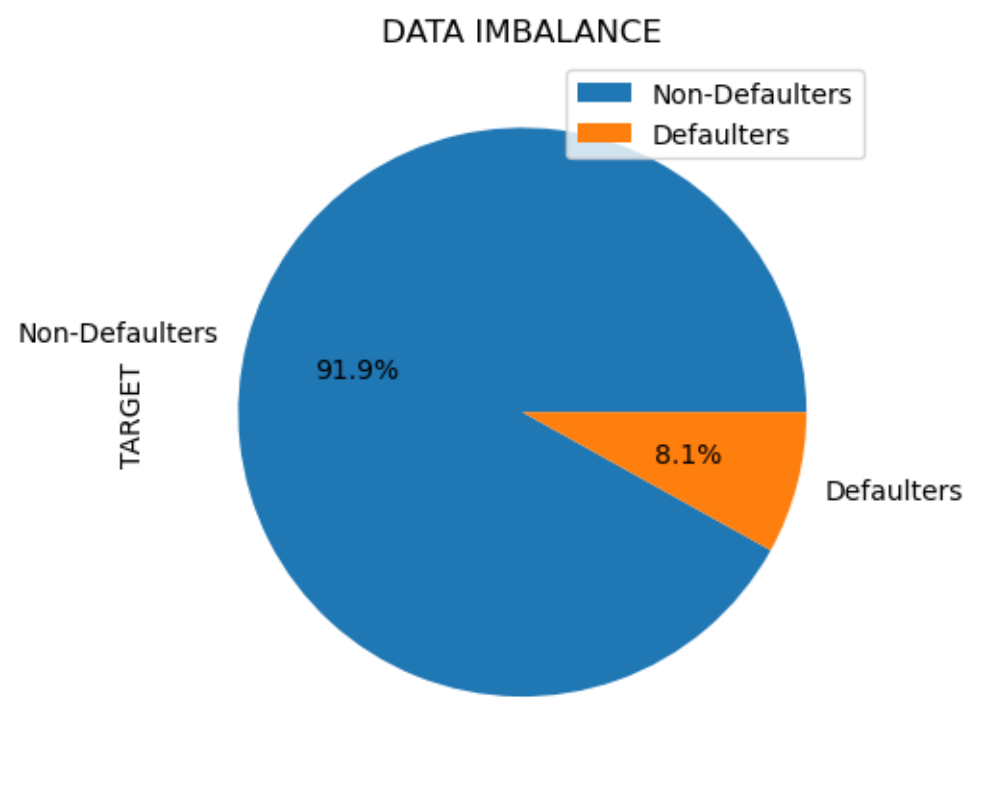


  - We can see a single value as an outlier.
  - 117000000.0 is the outlier value.

- Checking for outlier in **YEARS_EMPLOYED**
  - We can see an outlier value ~1000 which is an Outlier.



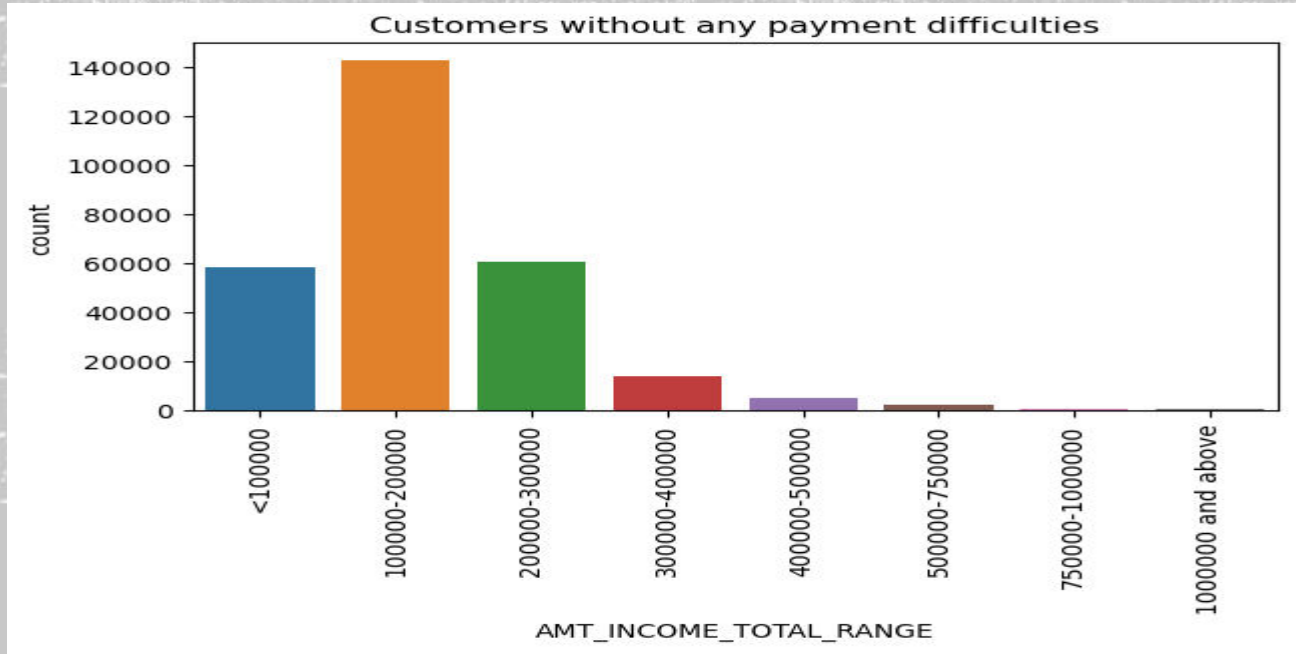
DATA IMBALANCE RATIO

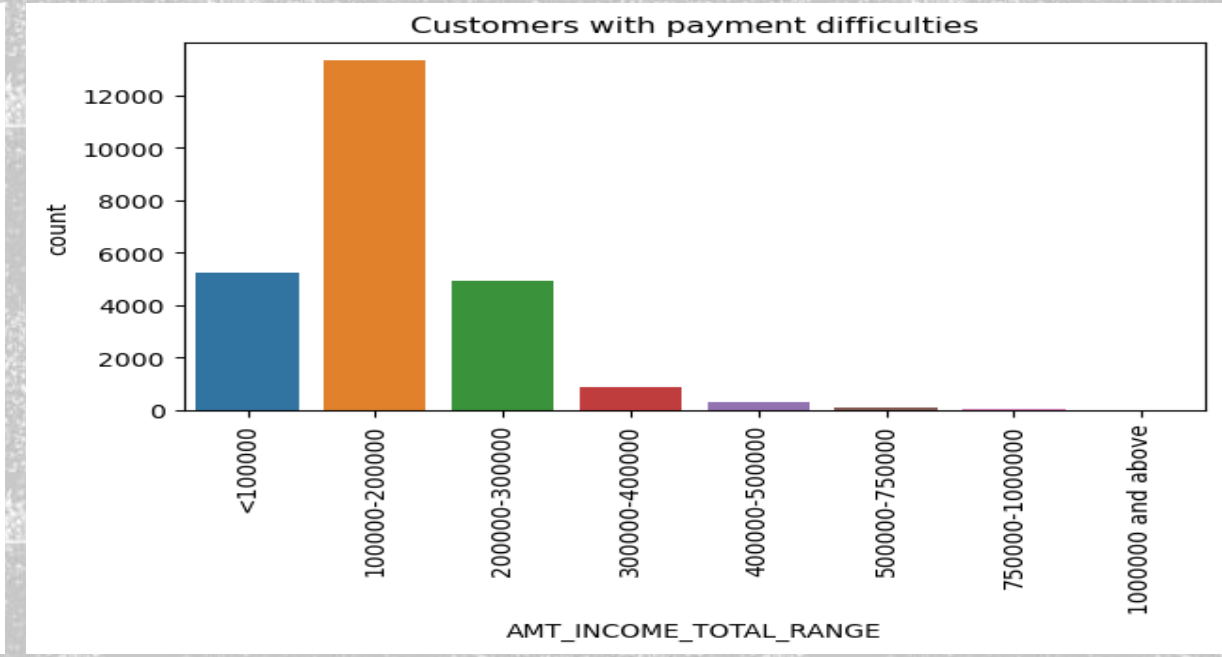


Data imbalance ratio between non-defaulters(0) and defaulters(1) is ~ 11:1

UNIVARIATE ANALYSIS

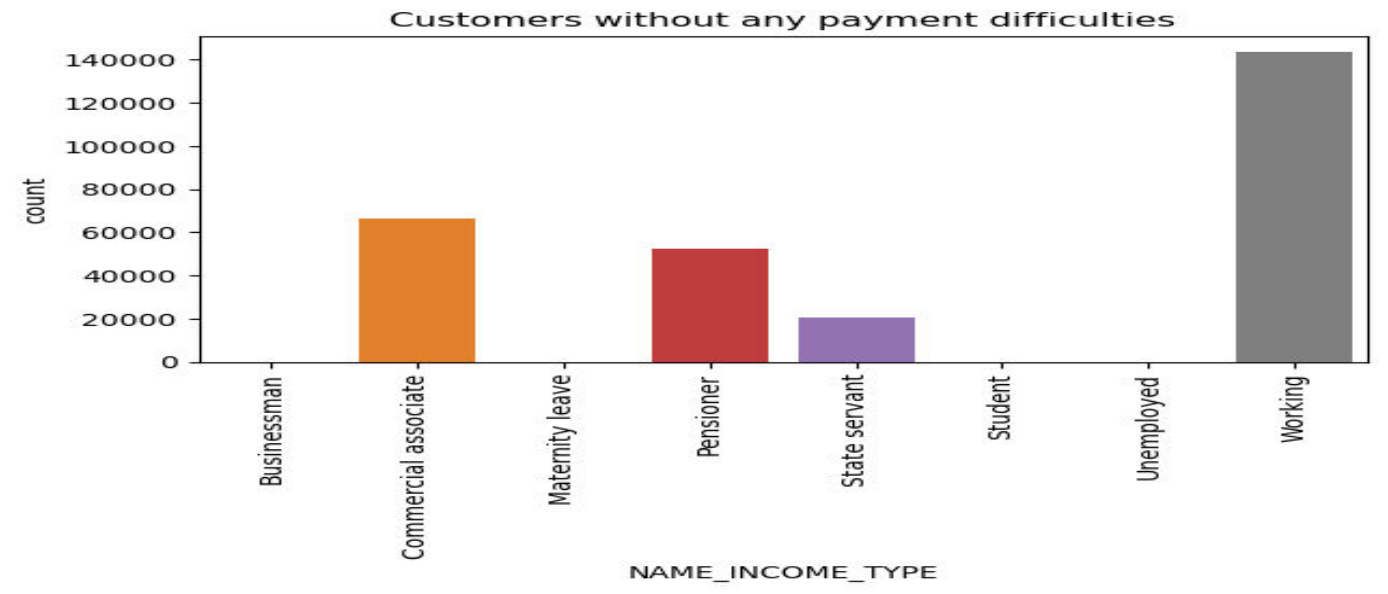
- Analysing **AMT_INCOME_TOTAL_RANGE**

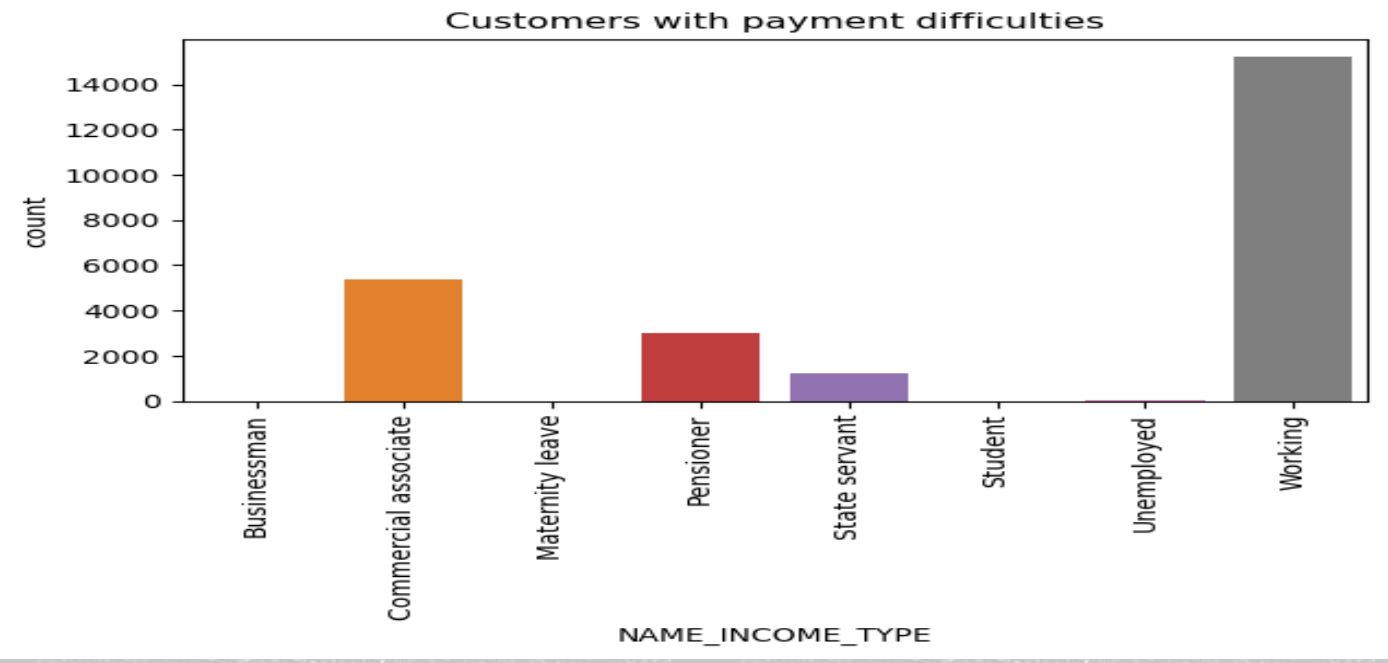




- There is a whole lot customers who are having income between 1Lakh and 2Lakhs without having any payment difficulties.
- Out of all the 250 customers having total income greater than 1M , only 13 are having difficulties to repay.

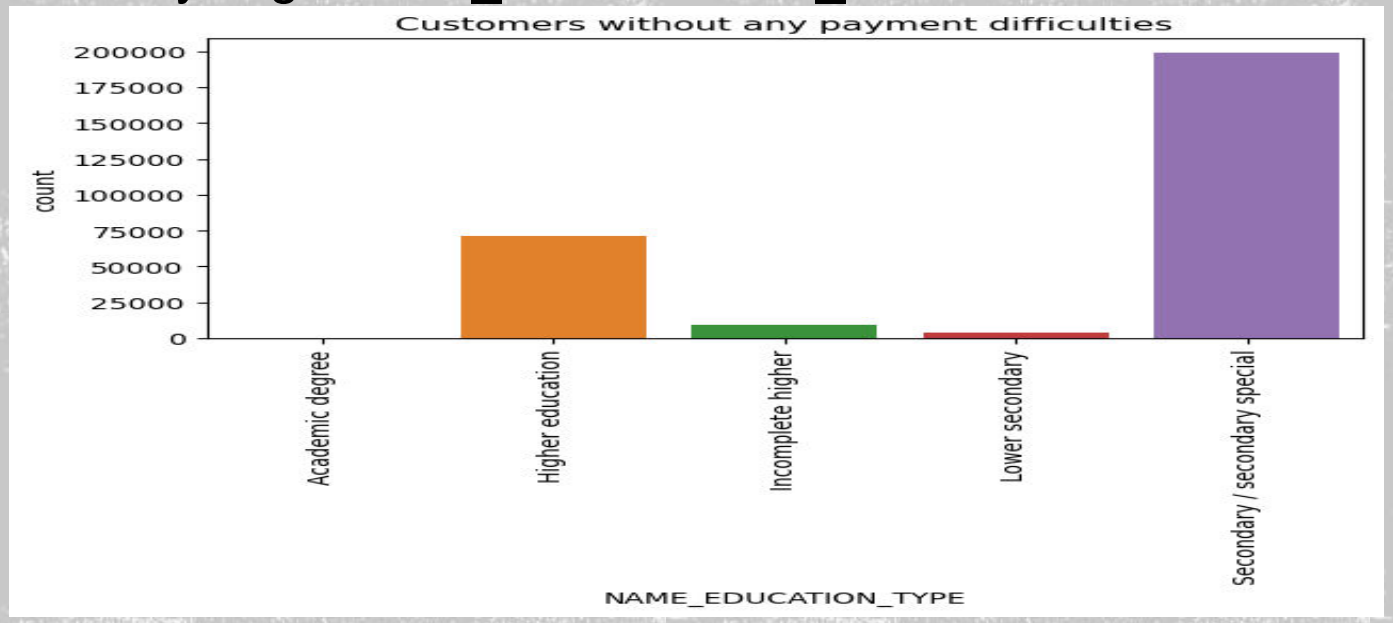
- Analysing **INCOME_TYPE**

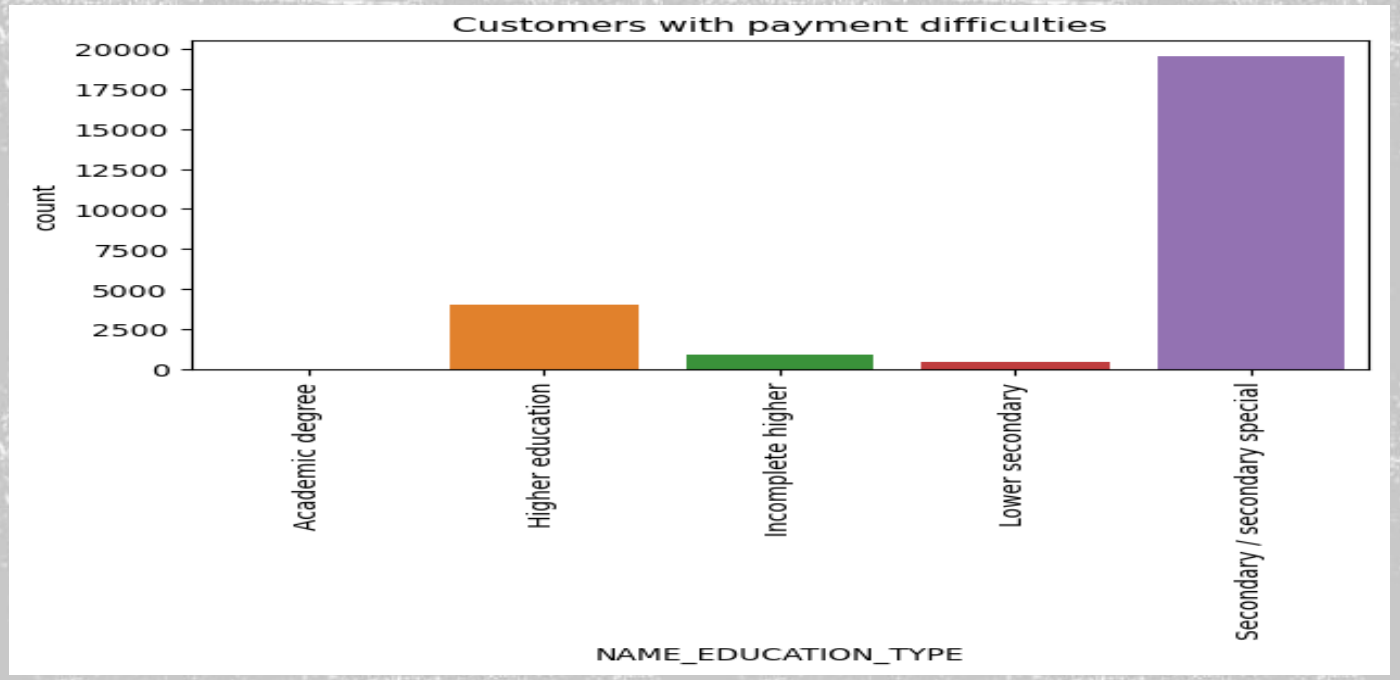




- Most of the customers who are taking loans are having INCOME TYPE as **working**.
- % of customers in **Working** category having difficulties in payment = **9.59** %.
- Customers who are having NAME_INCOME_TYPE as **Businessman** or **Student** is **not** having any **difficulties** when it comes to repay loans without difficulties, even if the no. of customers are very low in this case.

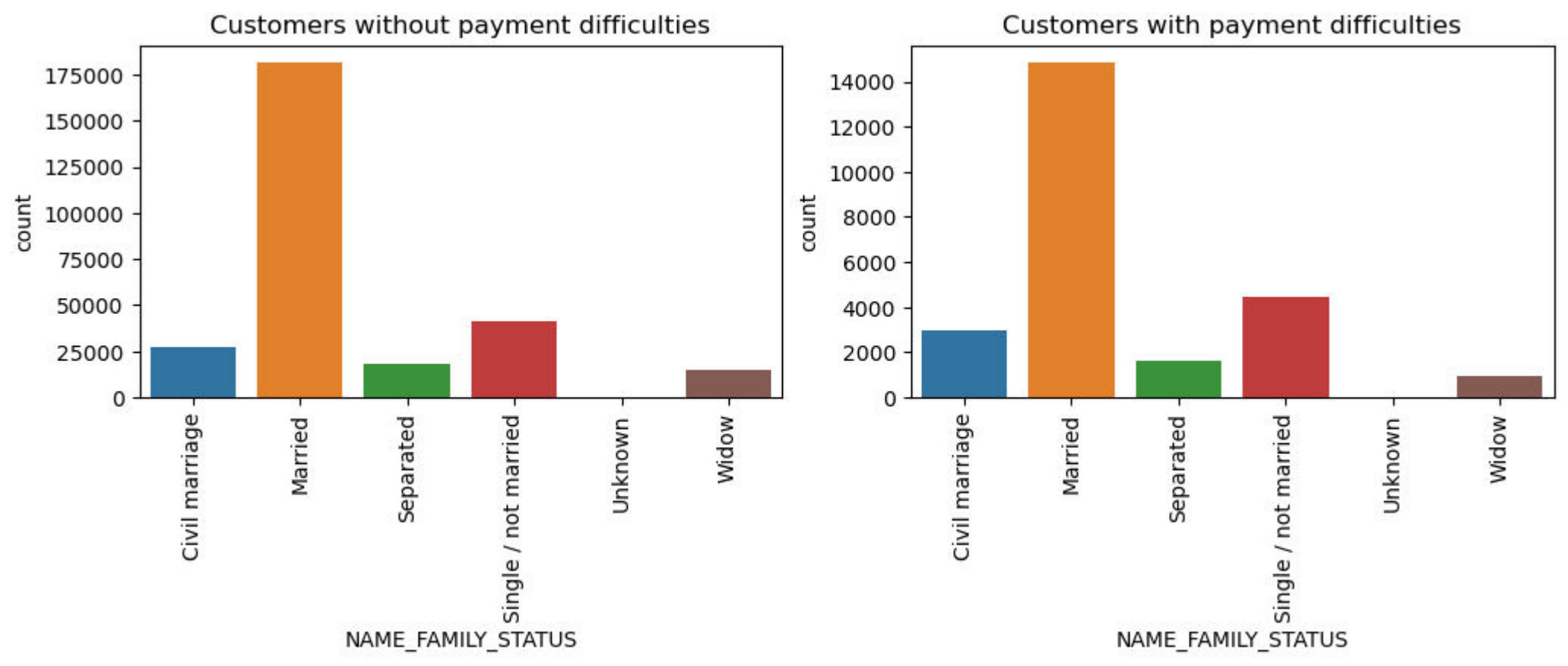
- Analysing **NAME_EDUCATION_TYPE**





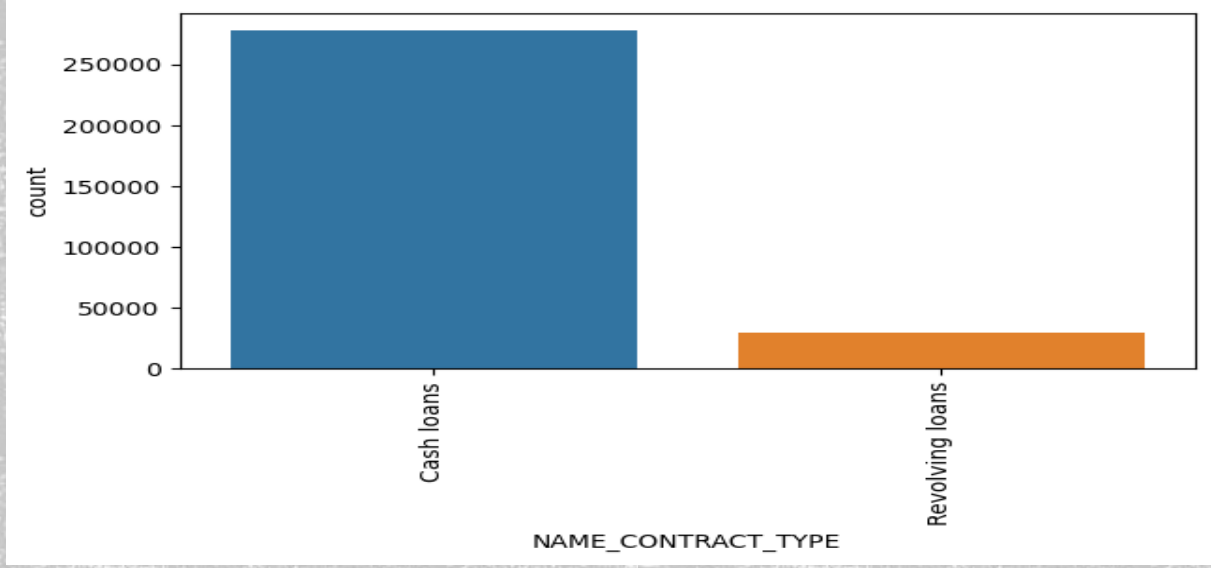
- Most of the customers are having **secondary/secondary special education** or **higher education.**

- Analysing **NAME_FAMILY_STATUS**



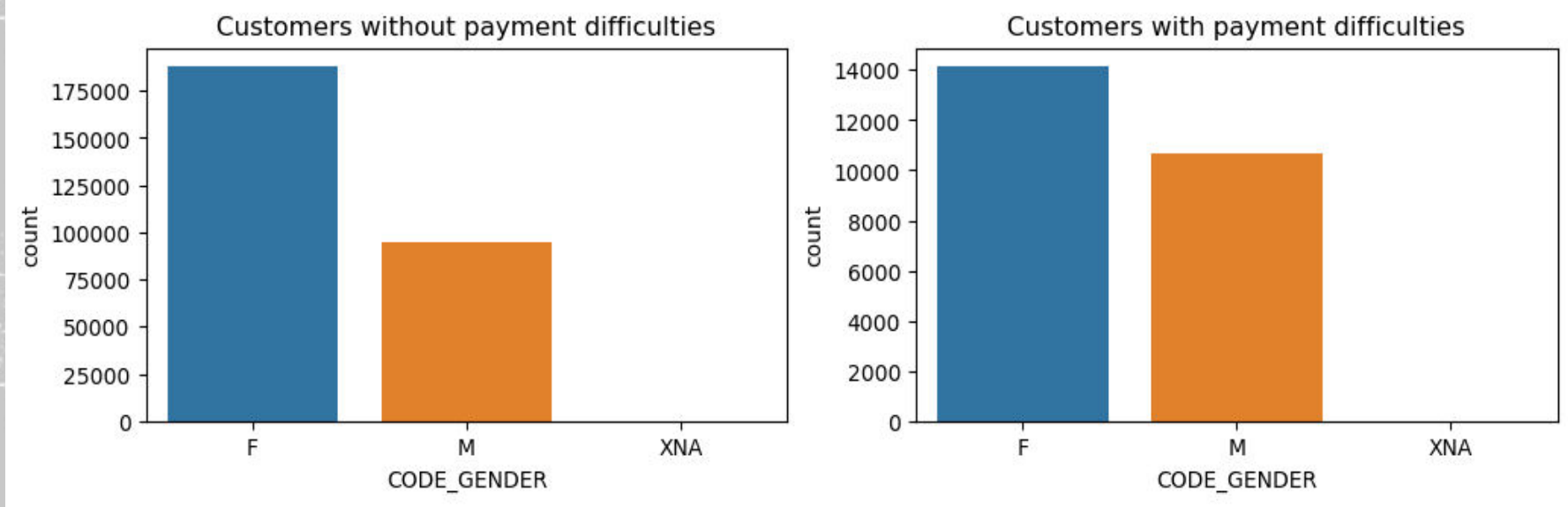
- Most of the customers are married.
- Seeing the statistics, it is always better to give out loans to Married because, we see that a total of around 64% of customers taking loans are married and out of these only 7.5% are facing difficulties repaying the loans.
- So, financially, it's a risk worth taking when giving out loans to Married customers.

- Analysing **NAME_CONTRACT_TYPE**



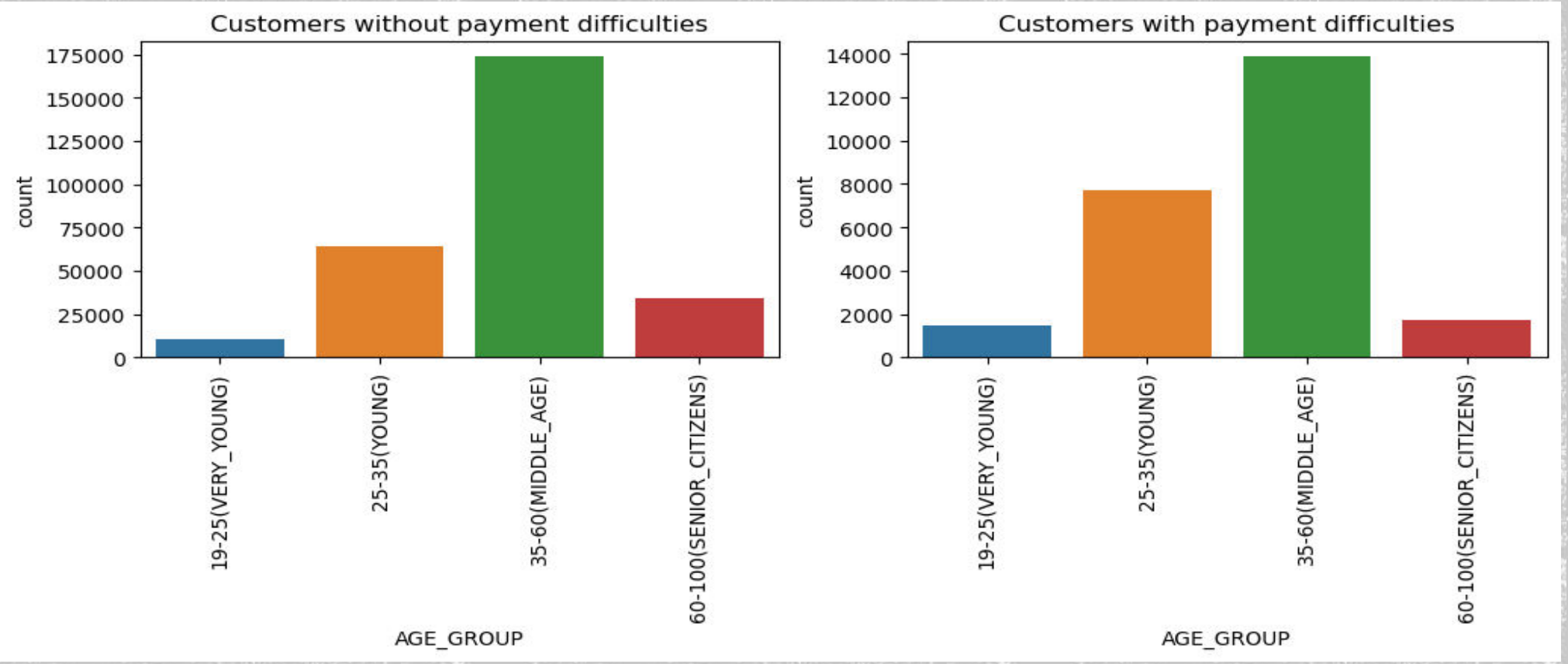
- **Most of the contract type are cash loans.**

- Analysing **CODE_GENDER**



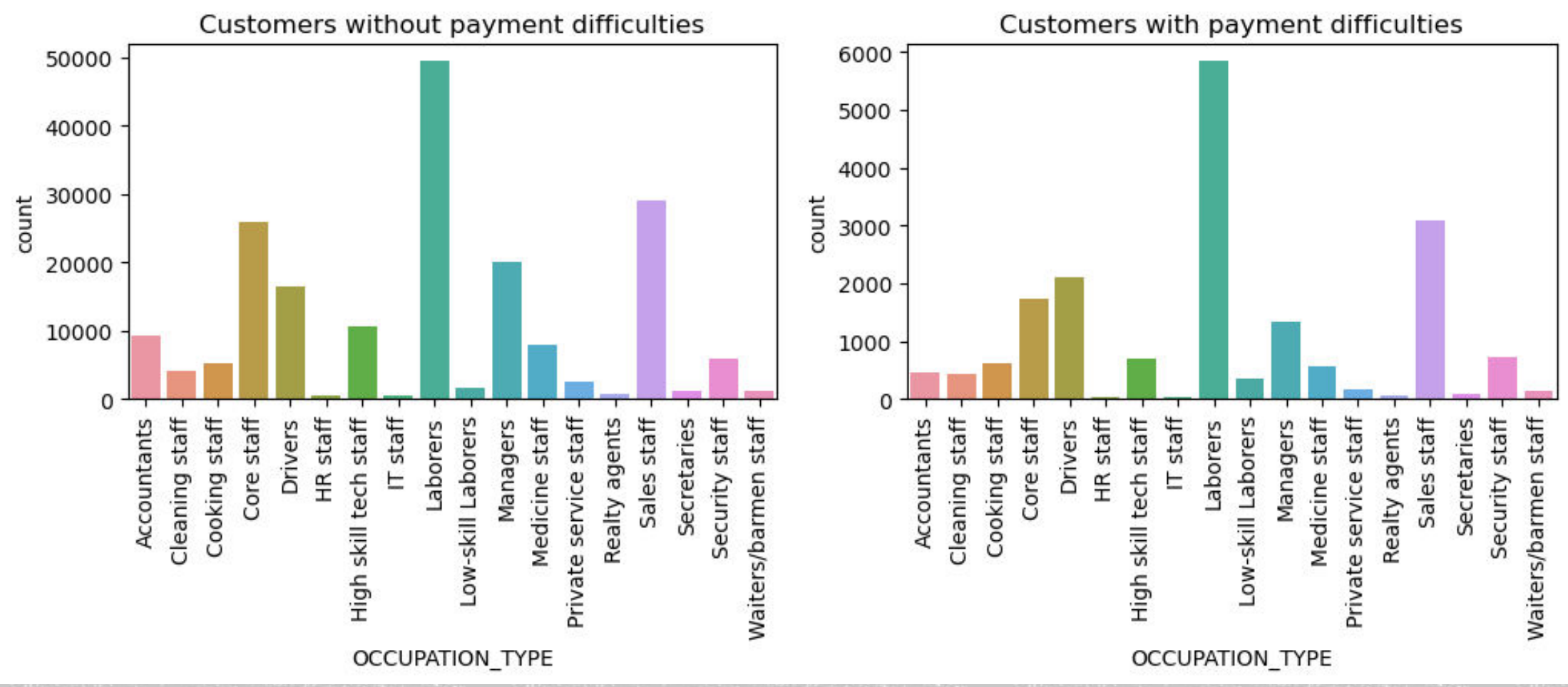
- Females are the most customers taking loans.
- Even if it seems that females are mostly likely to face difficulties in repaying loans while looking at the plot, it is actually reverse.
- We see that out of a total of **65.8%** of **female** customers, only below **7%** is having difficulties repaying while more than **10% of male** customers are **facing issues** repaying.

- Analysing **AGE_GROUP**



- Most of the customers are **middle aged people (35 - 60 years old)** who are applying for loans.
- Only around **7.3%** of middle age people are having **difficulties** paying loans back.

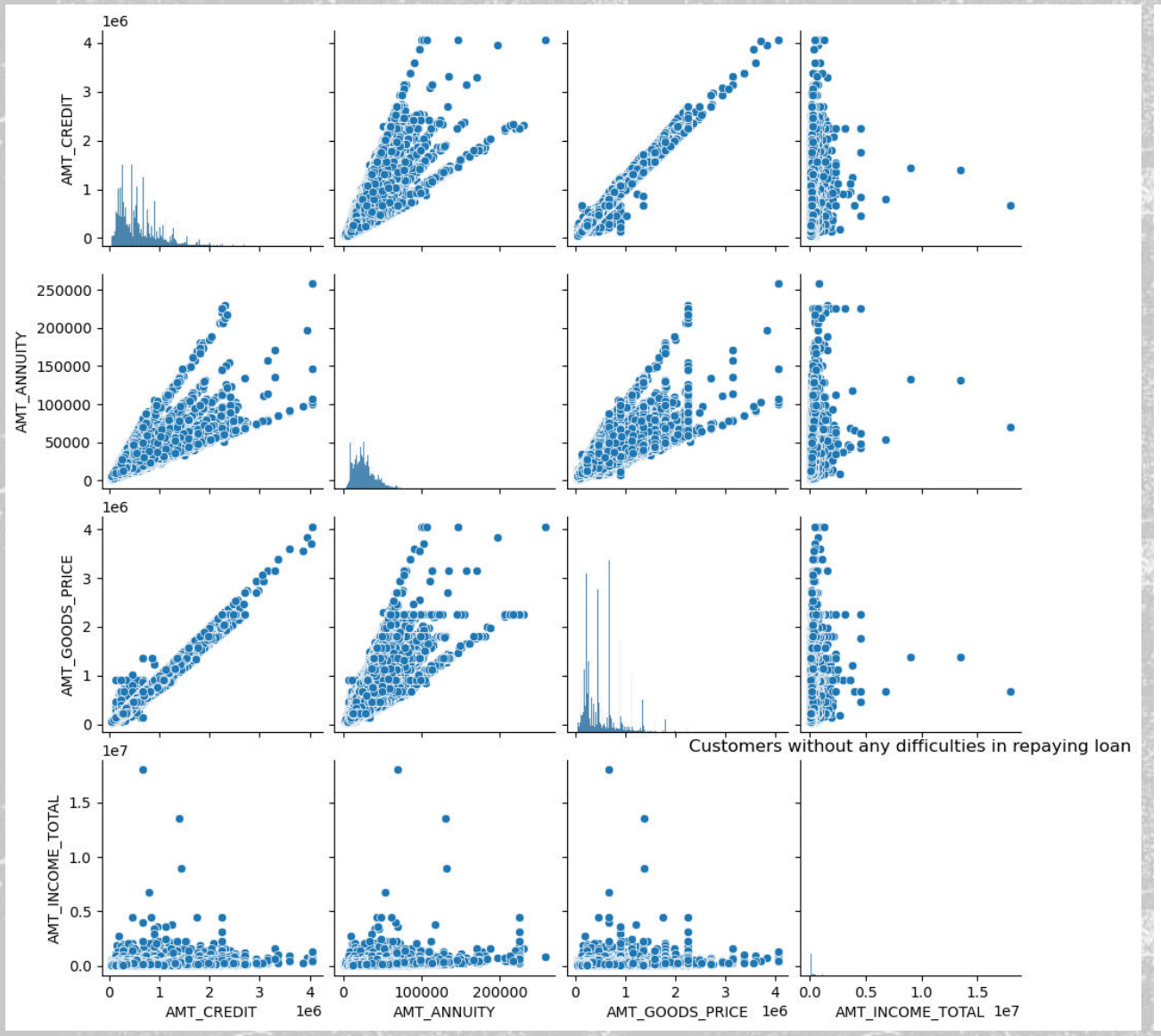
- Analysing **OCCUPATION_TYPE**

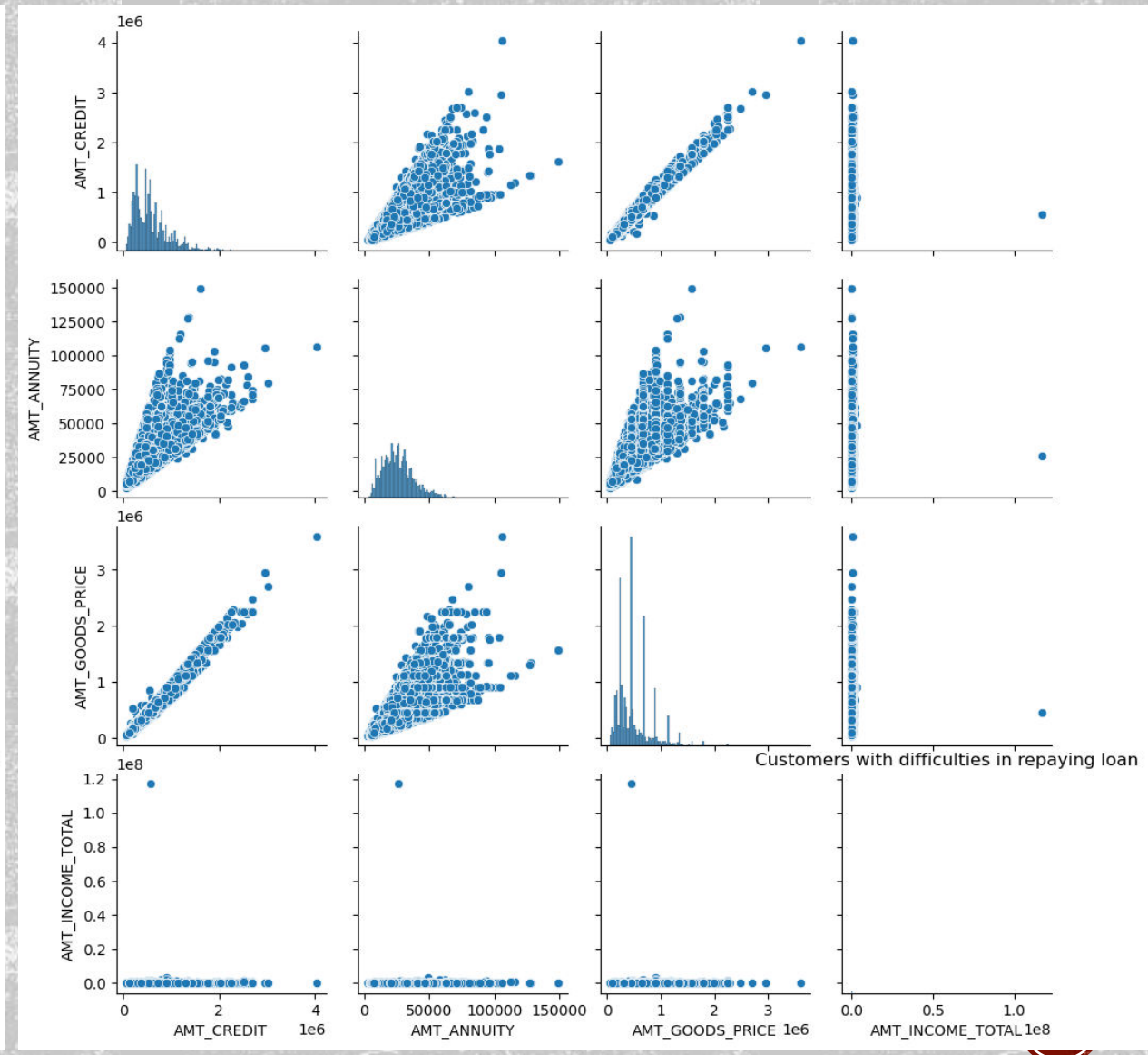


- Customers whose **OCCUPATION_TYPE** is **Laborer** are the ones who is **not facing** any issues while repaying and when it comes to the customers who are **facing difficulties** in repaying , we can see **Laborers topping** the count.
- We can see that more than **26%** of the whole customers taking loans are laborers.
- 10% of total customers as Laborers are having difficulties repaying loans.

BIVARIATE ANALYSIS - NUMERICAL - NUMERICAL ANALYSIS

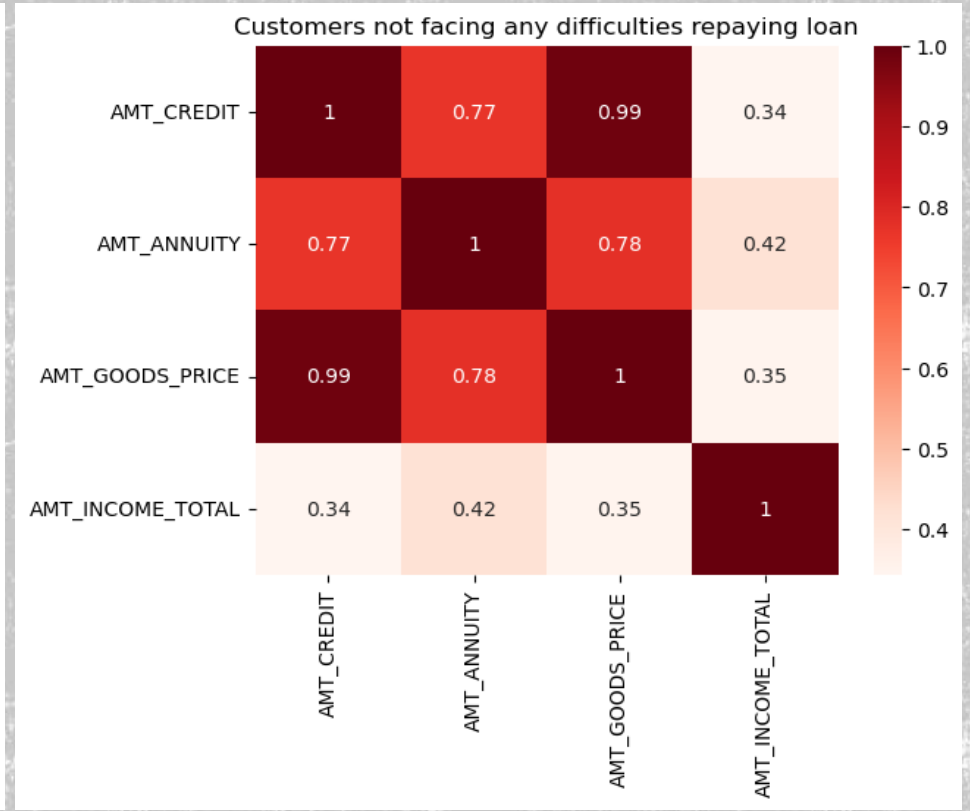
- ANALYSING **AMT_CREDIT, AMT_ANNUITY,AMT_GOODS_PRICE,AMT_INCOME_TOTAL**





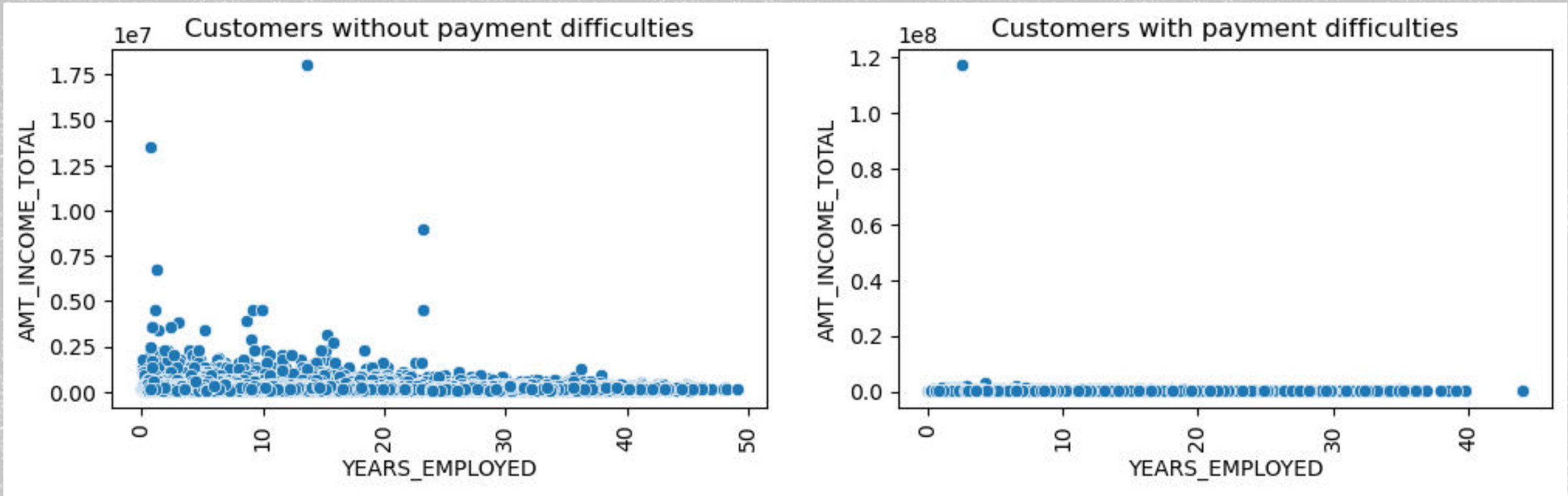
- **AMT_GOODS_PRICE increases** as **AMT_CREDIT increases** in cases where customers don't face difficulties in repaying as well as customers facing difficulties. So, we can guess that there is a **good positive correlation** between **AMT_GOODS_PRICE** and **AMT_CREDIT**.
- **AMT_ANNUITY increases** as **AMT_CREDIT increases**. This is because, when the **credit increases**, the **annuity also increases** which makes sense. Similarly, **AMT_ANNUITY increases** as **AMT_GOODS_PRICE increases** which can **lead to increase in credit** which is exactly what we are seeing in the plot.

- CORRELATION **AMT_CREDIT, AMT_ANNUITY,AMT_GOODS_PRICE,AMT_INCOME_TOTAL**



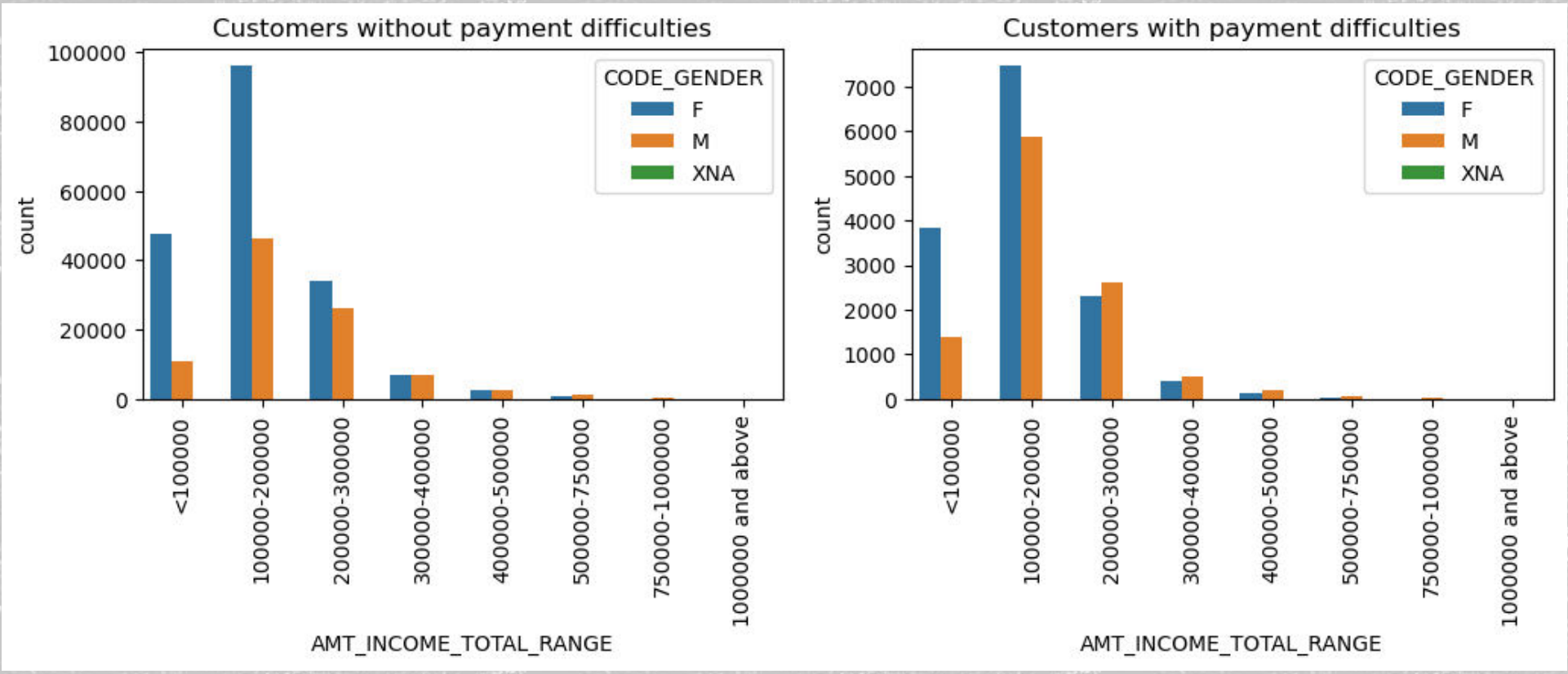
- From the heatmaps, we can understand that, correlation between **AMT_GOODS_PRICE** and **AMT_CREDIT** is the **highest**.

- Analysis between **AMT_INCOME_TOTAL** and **YEARS_EMPLOYED**



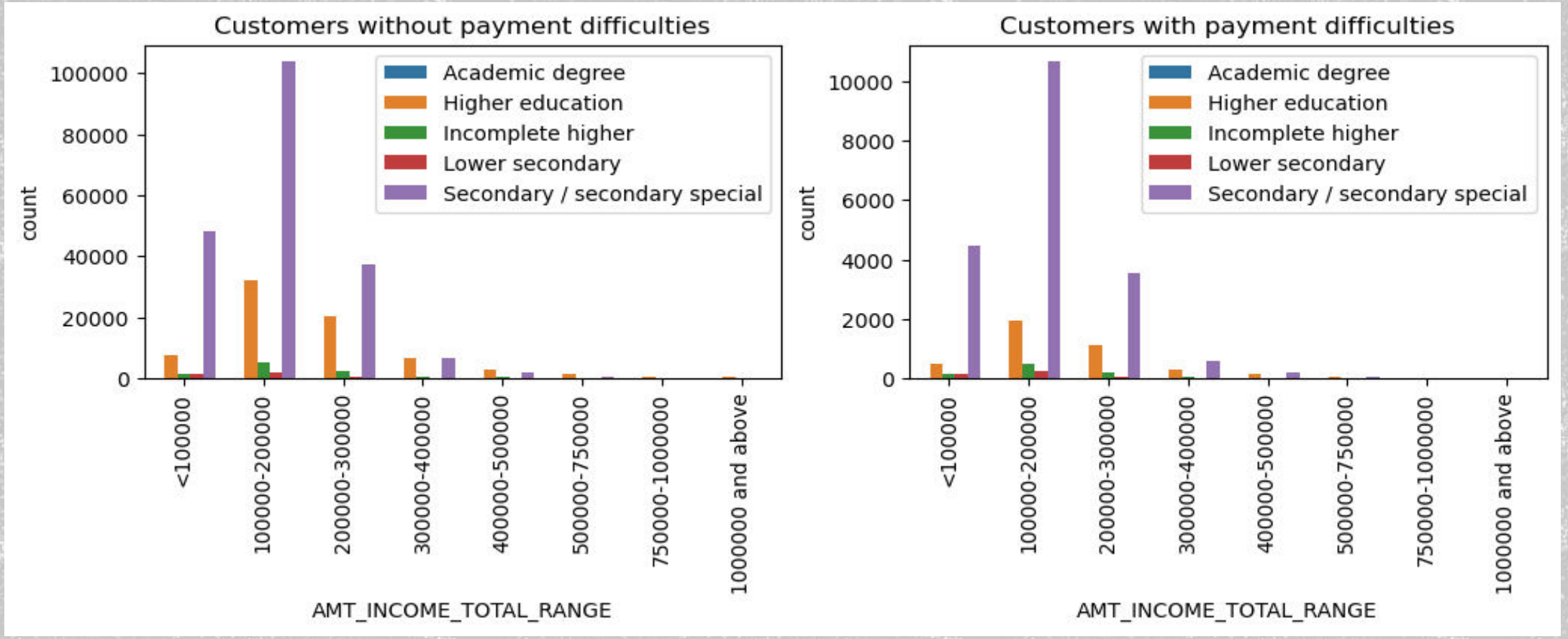
- Personally, I was **expecting** to see an **increase** in the **AMT_INCOME_TOTAL as YEARS_EMPLOYED increases**. But **can't see it really happening** in this plot.

- **CATEGORICAL - CATEGORICAL ANALYSIS**
- Analysing **AMT_INCOME_TOTAL_RANGE** and **CODE_GENDER**



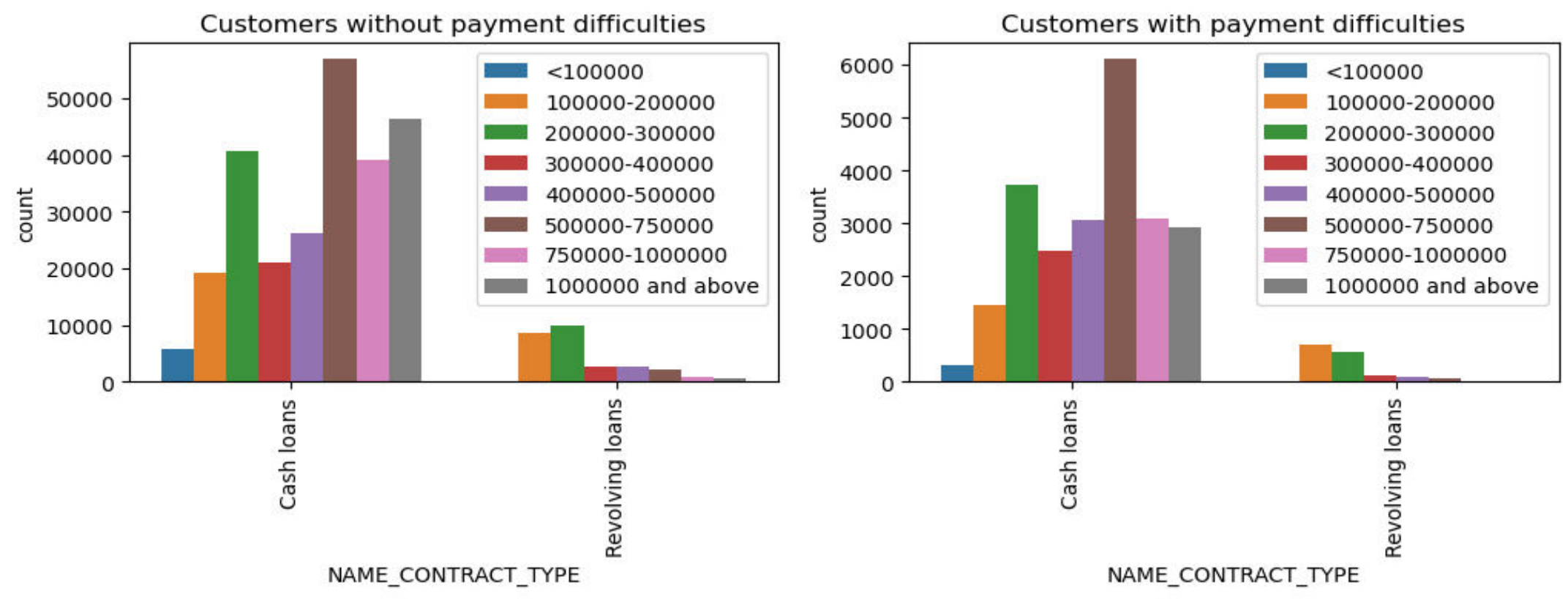
- In most of the INCOME_RANGE, **women are finding it easy to repay loan** without difficulties than men.
- When it comes it defaulters, **after the income range of 200K**, **men are finding it difficult to repay loans than women.**

- Analysing **AMT_INCOME_TOTAL_RANGE** and **NAME_EDUCATION_TYPE**



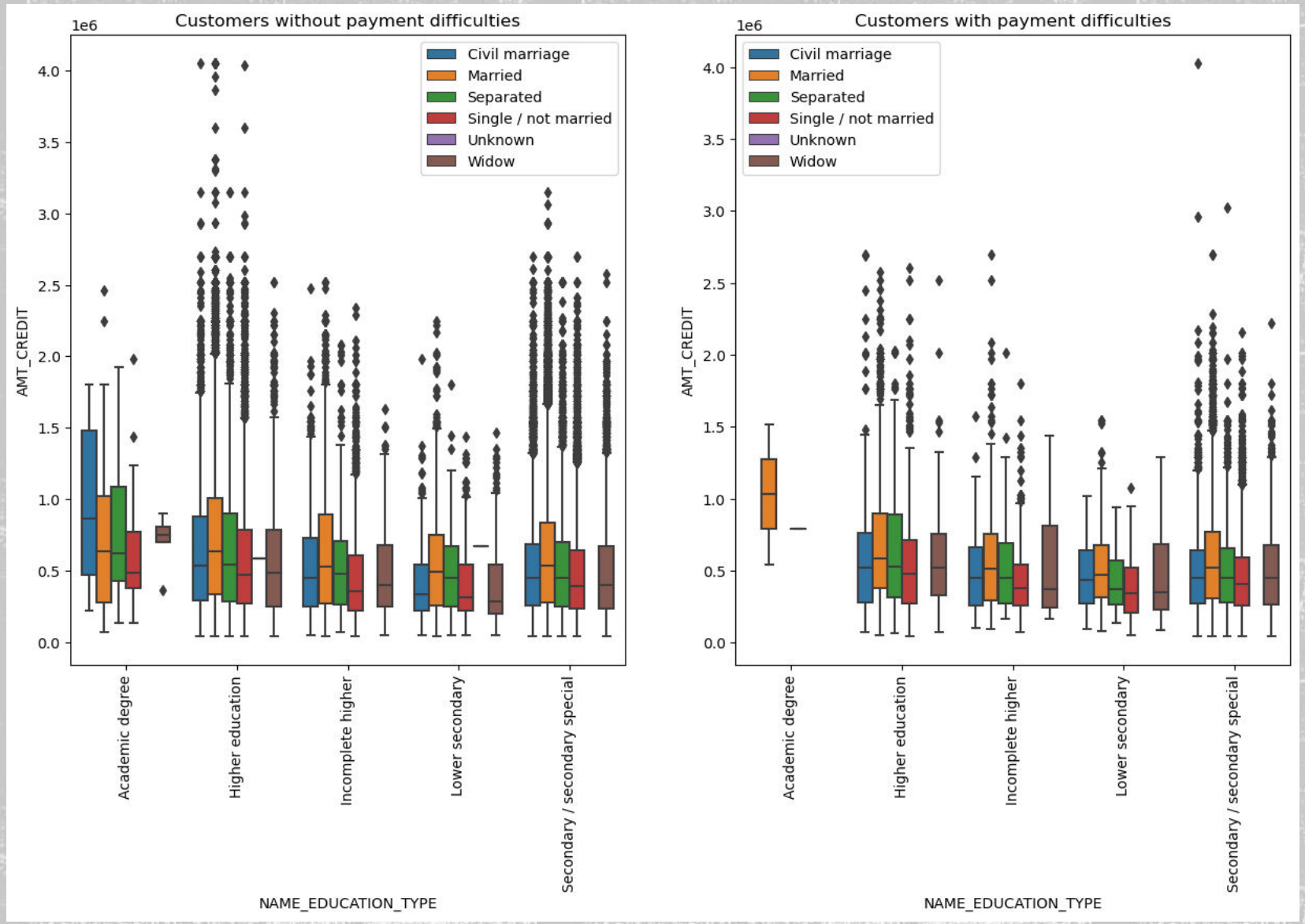
- **Secondary special education** has **higher income** in all ranges but also **tops** the category where customers **finds it difficult to repay** loan and also **tops repay loans** on time.

- Analysing **NAME_CONTRACT_TYPE** and **AMT_CREDIT_RANGE**



- In both the cases, the **credit range is highest** for the range **between 5Lakhs to 7.5 lakhs** and most of the customers are choosing **Cash loans** rather than revolving loans.

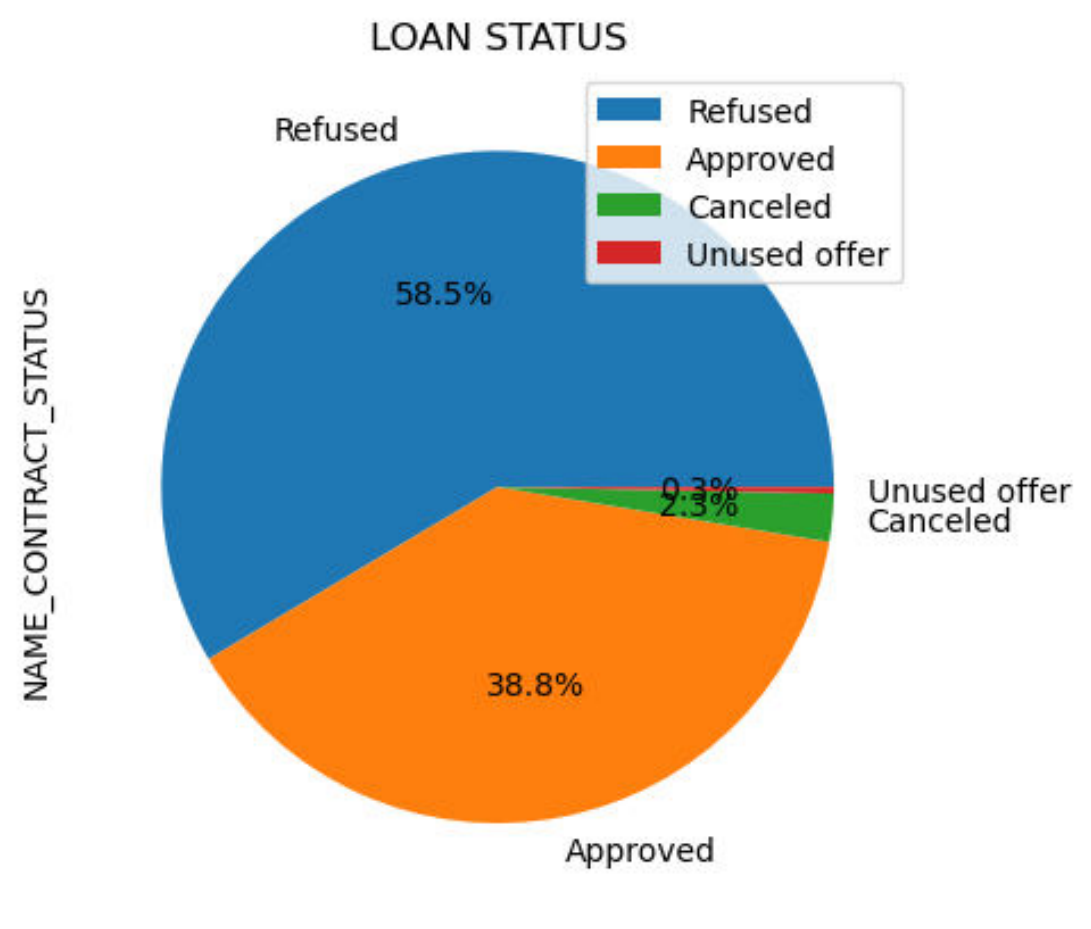
Analysing **AMT_CREDIT** and **NAME_EDUCATION_TYPE**



- In most education type ,**married customers** are finding it difficult to repay the loans without difficulties.
- The **median** of married customers is **high** compared to other categories in most education type whether finding it difficult to repay / not.
- Customers who are **single** and are having **Lower secondary** are getting the lowest **median** of **AMT_CREDIT**.
- **Higher education** customers gets **more AMT_CREDIT** than other education type customers.
- **Higher education category** has the **highest of outliers** in **customers without difficulty** while **secondary special category** has the **highest of outliers in customers with difficulty.**

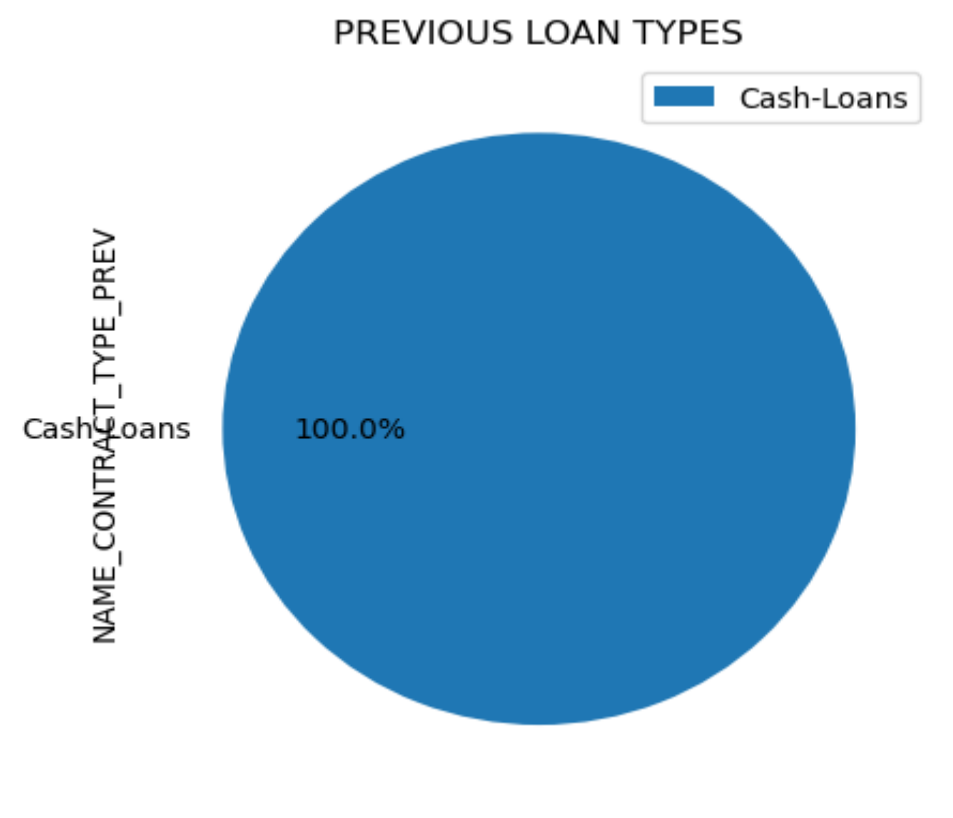
Analysis after merging-*'application_data.csv' and 'previous_application.csv'*

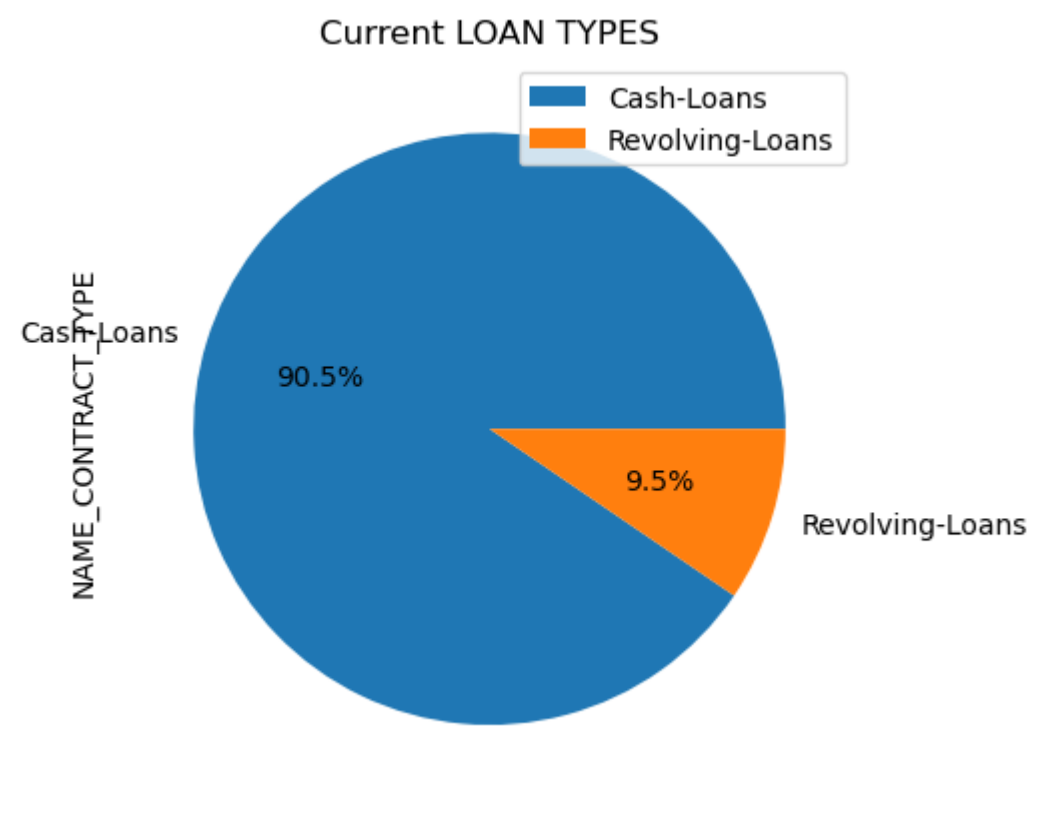
- **UNIVARIATE ANALYSIS**
  - Analysis of **NAME_CONTRACT_STATUS**



- **Most** of the contracts are **refused** when compared to accepted or other status.
- **Unused** offers are less than **0.5%** which indicated good campaign of loans and customers are aware of loans.
- **Less than 3%** of customers **canceled** the loans.

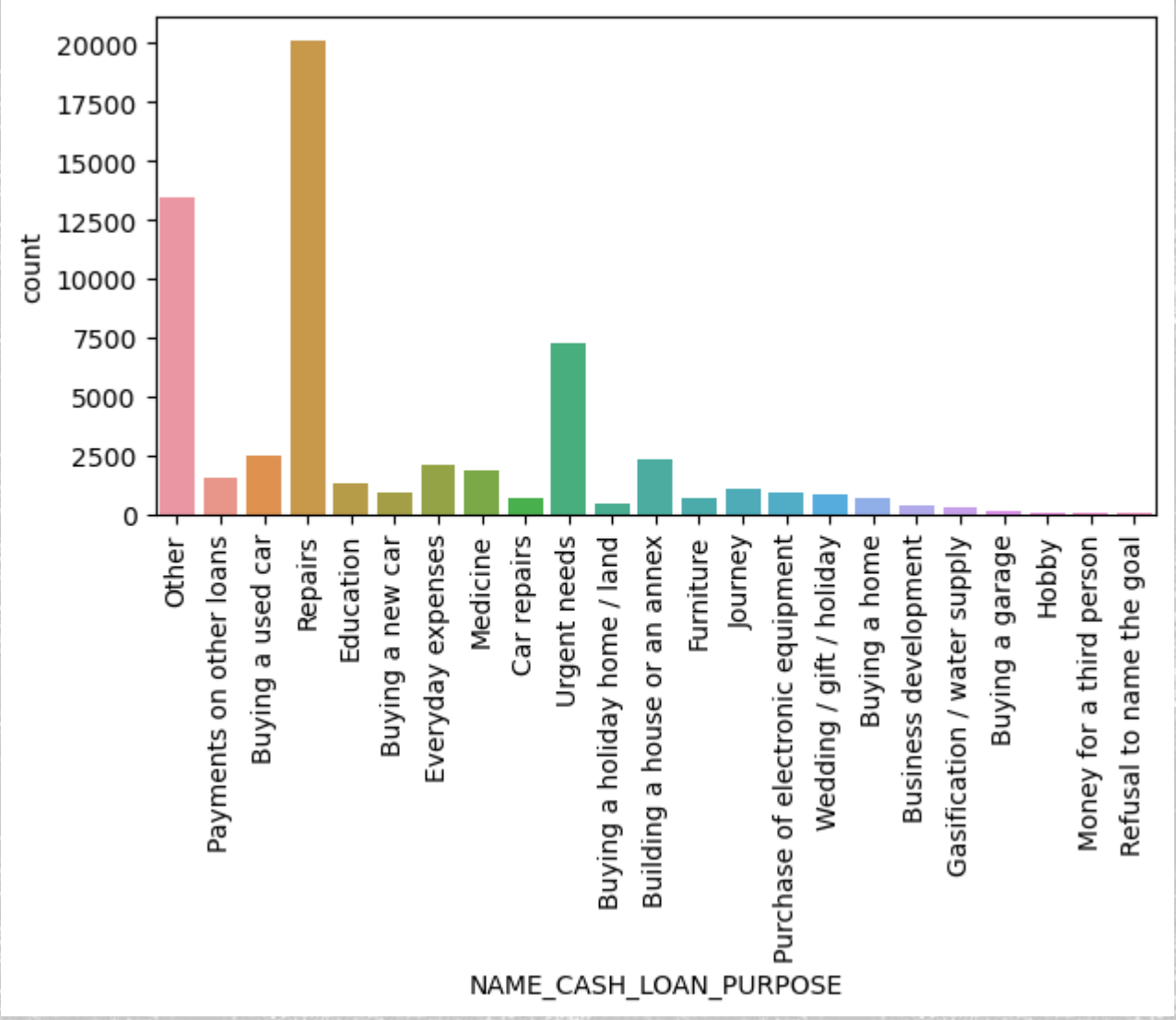
- Analysis on **NAME_CONTRACT_TYPE** (in application_data.csv and previous_application.csv)





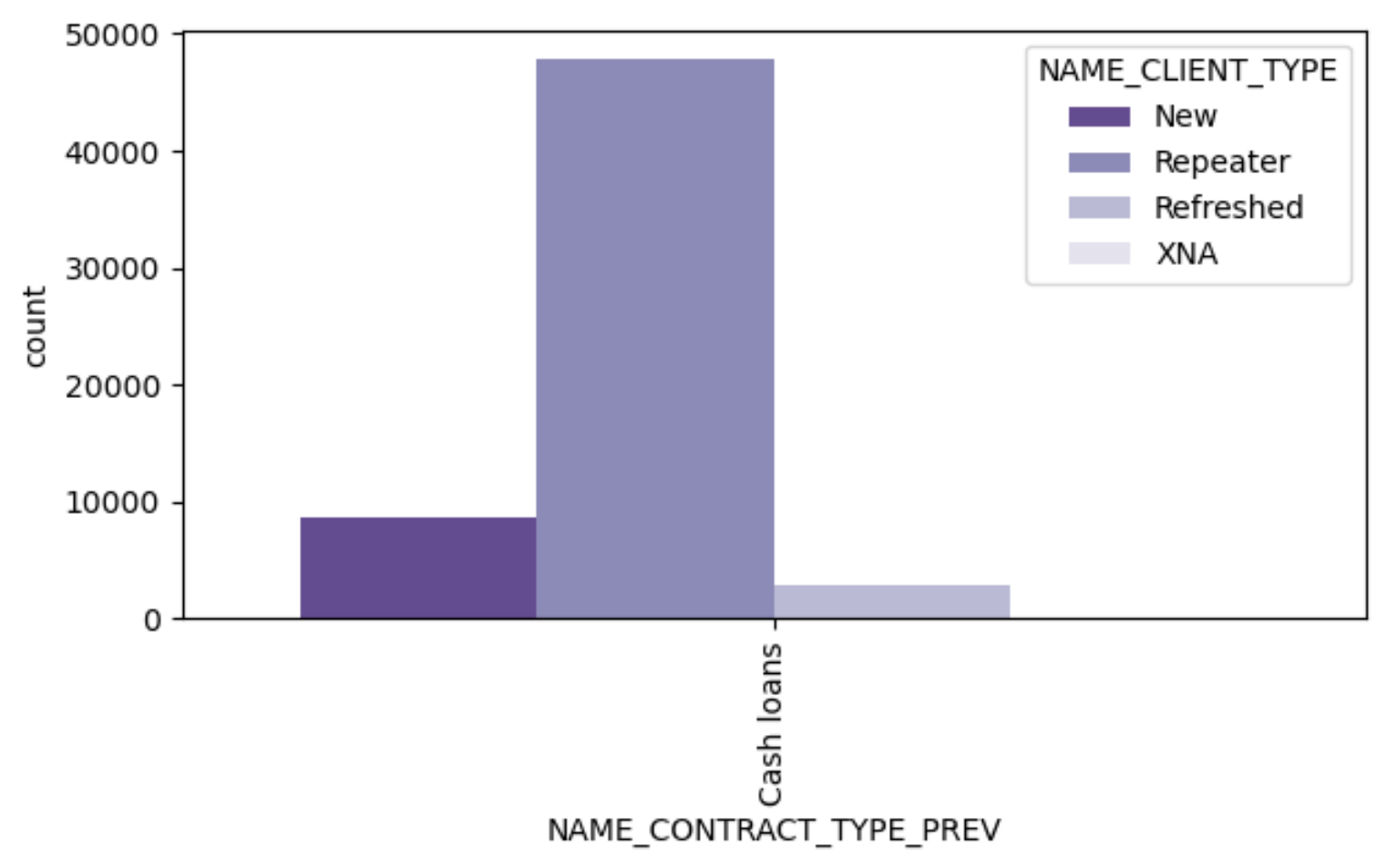
- From the above charts, we can understand that there is an **increase in the revolving loans by ~9.5%** when compared to previous data.

- Analysis on **CASH_LOAN_PURPOSE**



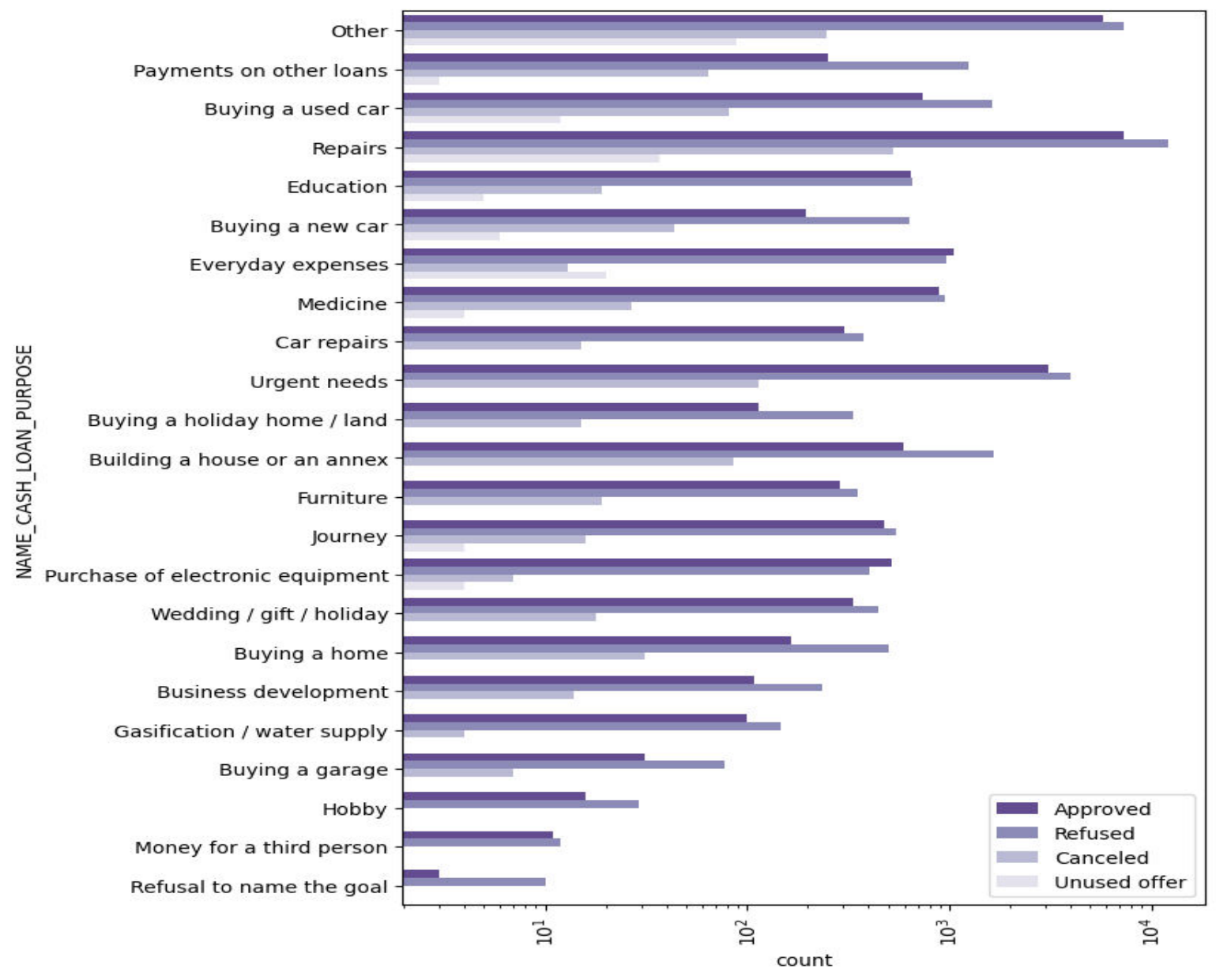
- **Most of the customers are taking loans for Repairs or Urgent needs.**

- ANALYSING **NAME_CONTRACT_TYPE_PREV** and **NAME_CLIENT_TYPE**



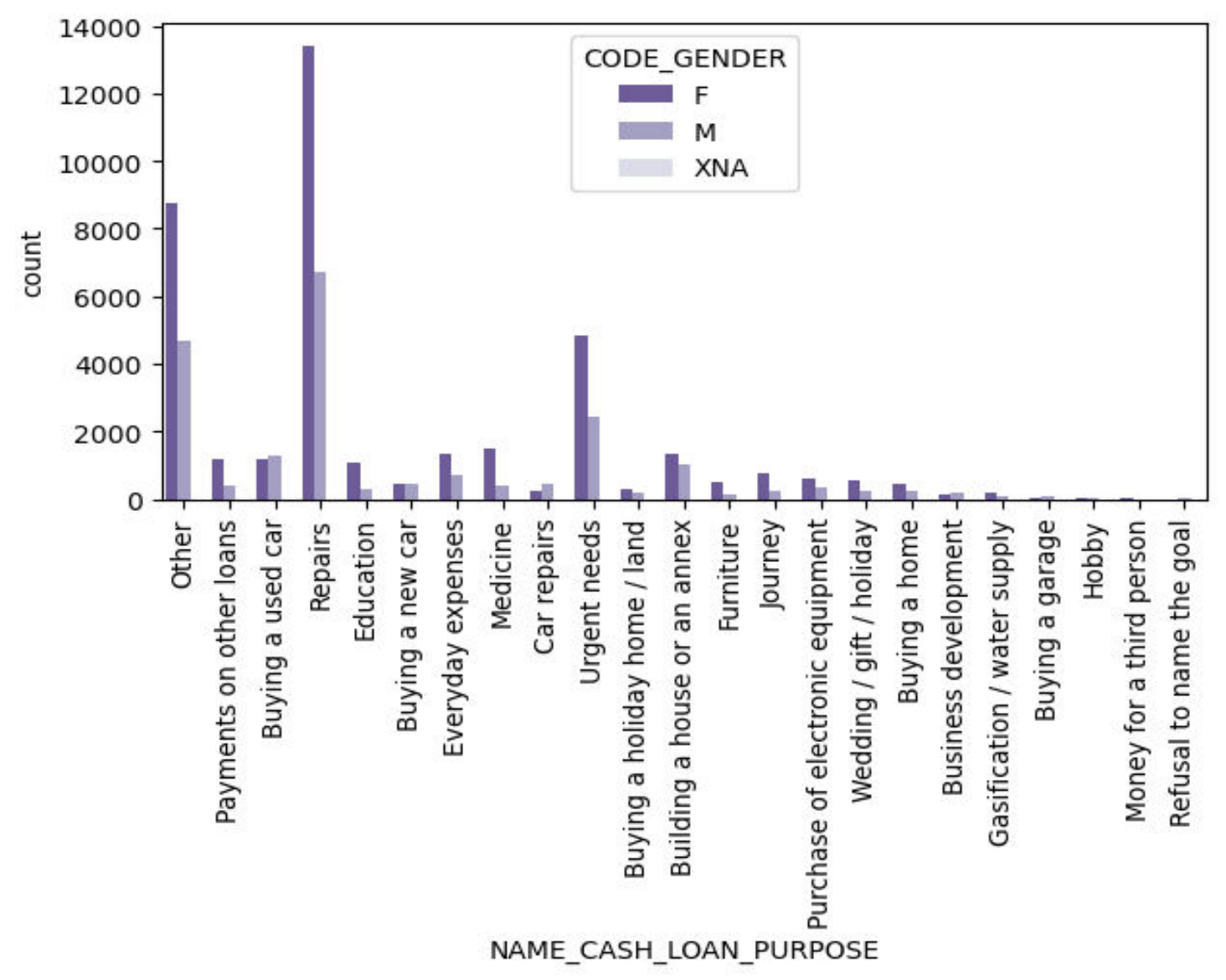
- All the previous data has contract type as **loans**.
- Most of the **customers are Repeaters**.
- There are nearly around **10K new appliers.**

ANALYSING **NAME_CASH_LOAN_PURPOSE** and **NAME_CONTRACT_STATUS**



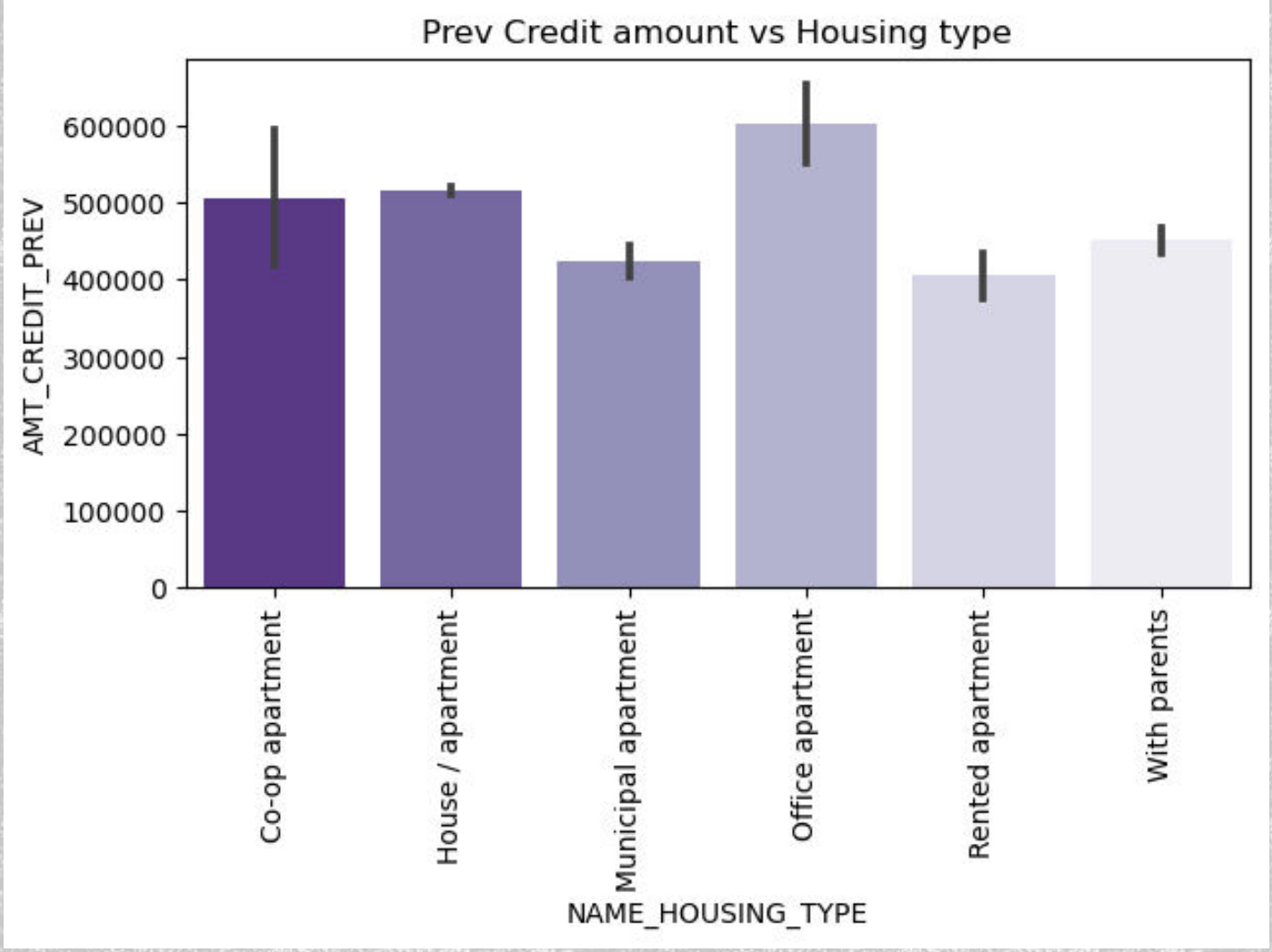
- **Approved status** is **greater** than **Refused** only when loans are taken **for everyday expenses** and **purchase of electronic equipment.**
- When it comes to loans for **payment for other loans**, the **rejection rate is very high** compared to its approved rate.
- **Rejection rate is highest for repairs, urgent needs and building house or annex.**
- For **education purposes**, the **approved** and **refused** status is barely **equal**.
- Most of the **customers have canceled** their plan to take loans when its **for urgent needs or for building a house**.
- This gives us insights about the housing loans the bank is providing. Maybe its not likely to satisfy the needs of customer when it comes to housing loans.

ANALYSING **NAME_CASH_LOAN_PURPOSE and CODE_GENDER**



- **Most of the loans are being taken for Repairs or for Urgent needs and by females.**

- ANALYSING **AMT_CREDIT_PREV** and **NAME_HOUSING_TYPE**

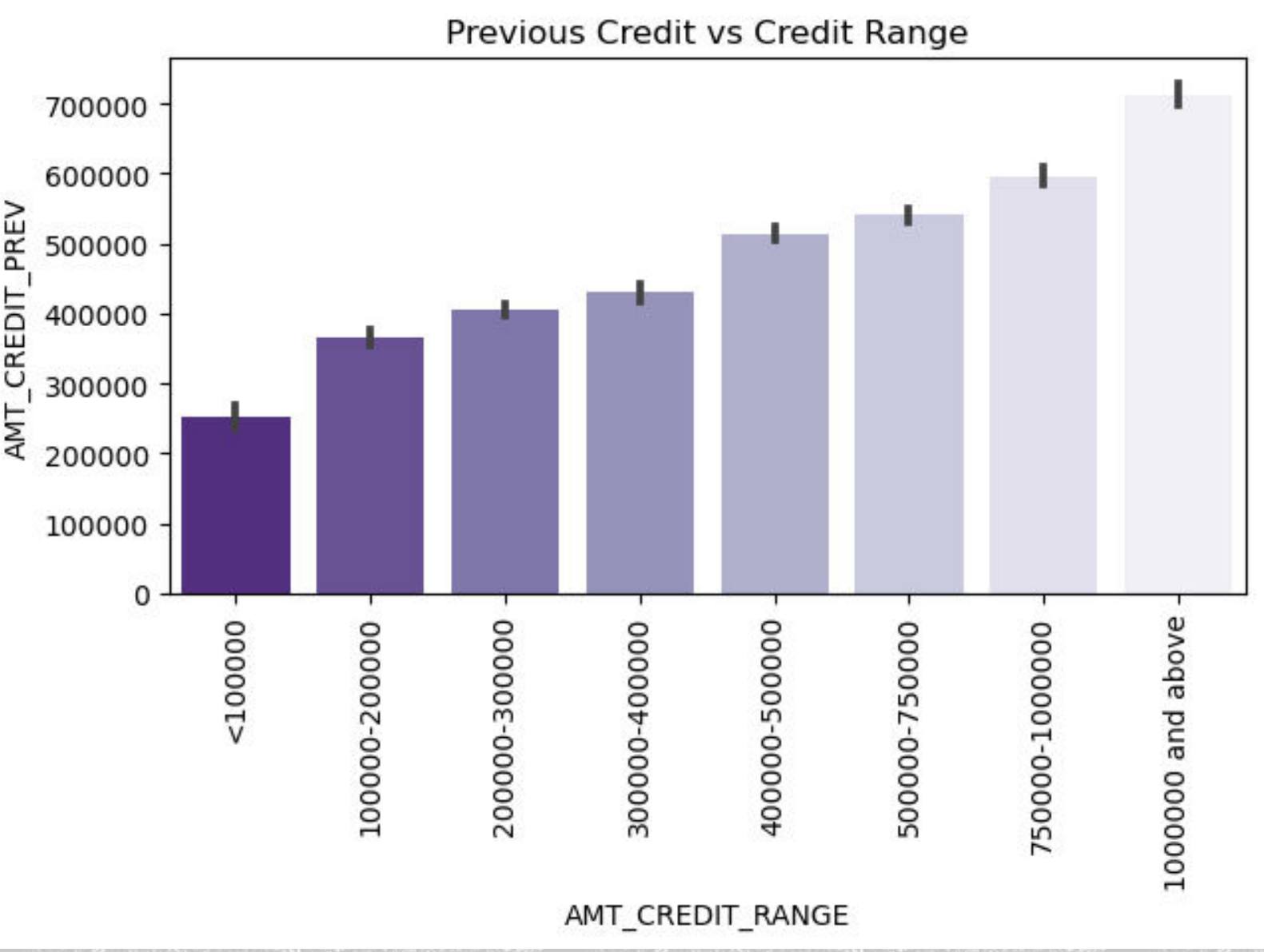


- **Credit amount (prev) is highest for customers with office apartment, Co-op apartment and House.**

ANALYSING **AMT_CREDIT_PREV , TARGET** and **NAME_HOUSING_TYPE**

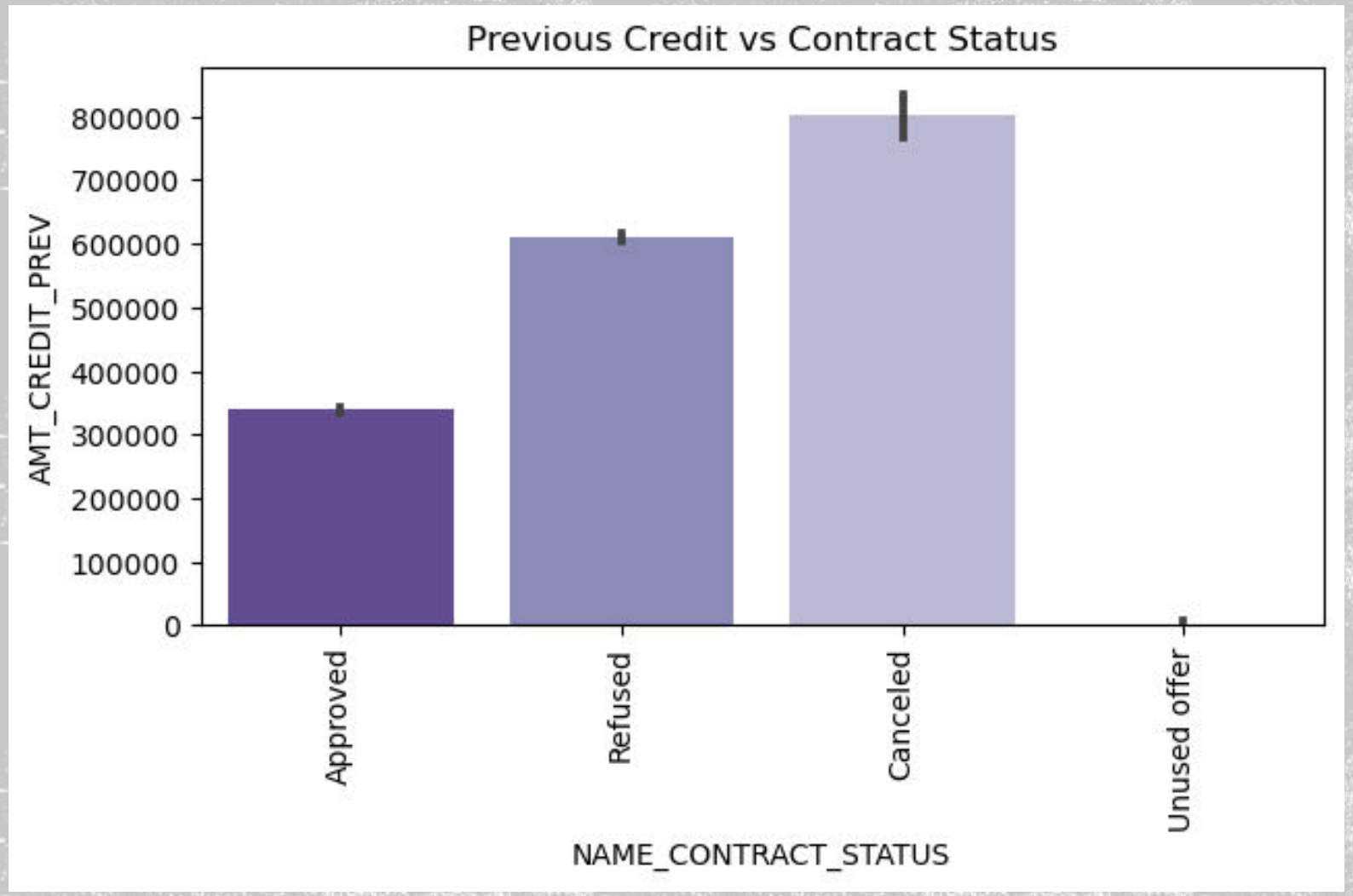
- **Defaulters** are **maximum** when their **housing type is co-op apartment** and when their **previous credit ranges between 500k and 600k**.
- **Defaulters** are **significantly low compared to its Non-defaulters** when customers are in **rented apartments**.
- **Credit amount is maximum** for customers in **co-op apartments**.

ANALYSING **AMT_CREDIT_PREV** and **AMT_CREDIT_RANGE**



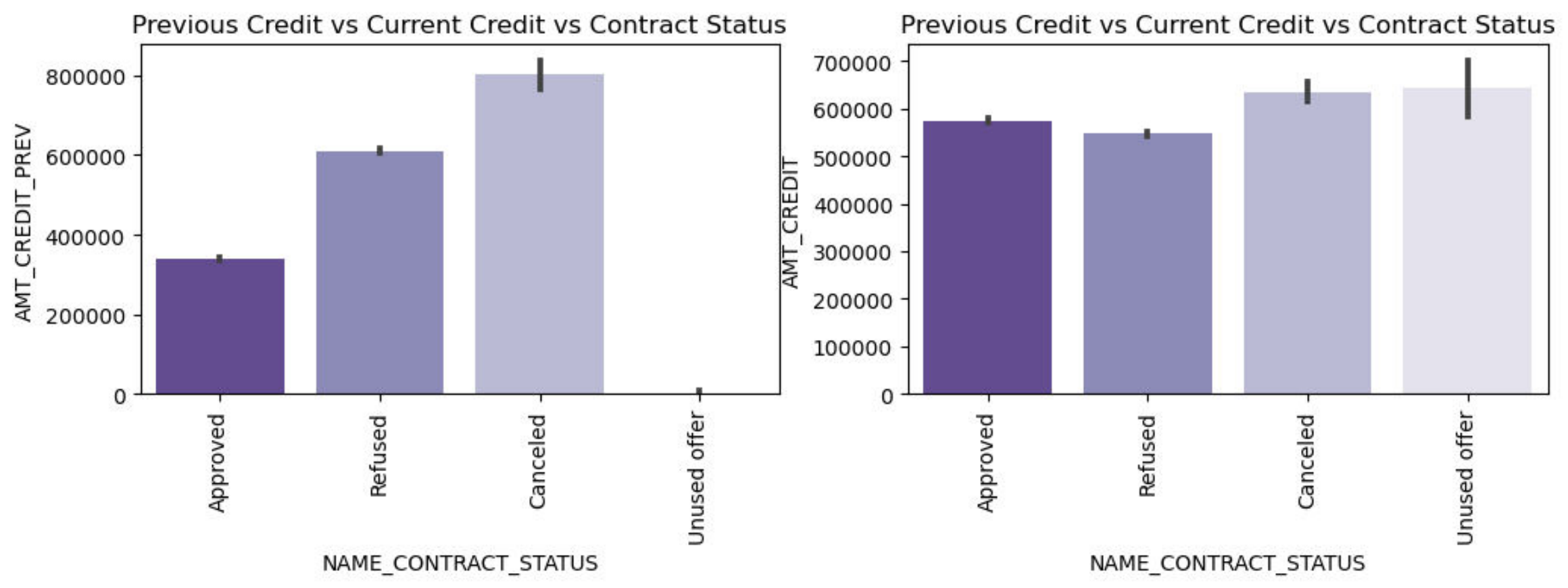
- Comparing to previous credit amount, the **current credit amount is significantly decreased for ranges below 100k , and ranges between 100k to 300k**. And there is a **slight decrease for ranges between 300k - 500K**.
- Comparing to previous credit amount, the **current credit amount is significantly INCREASED ranges between 500k to 300k and 1M and above**.

ANALYSING **AMT_CREDIT_PREV** and **NAME_CONTRACT_STATUS**



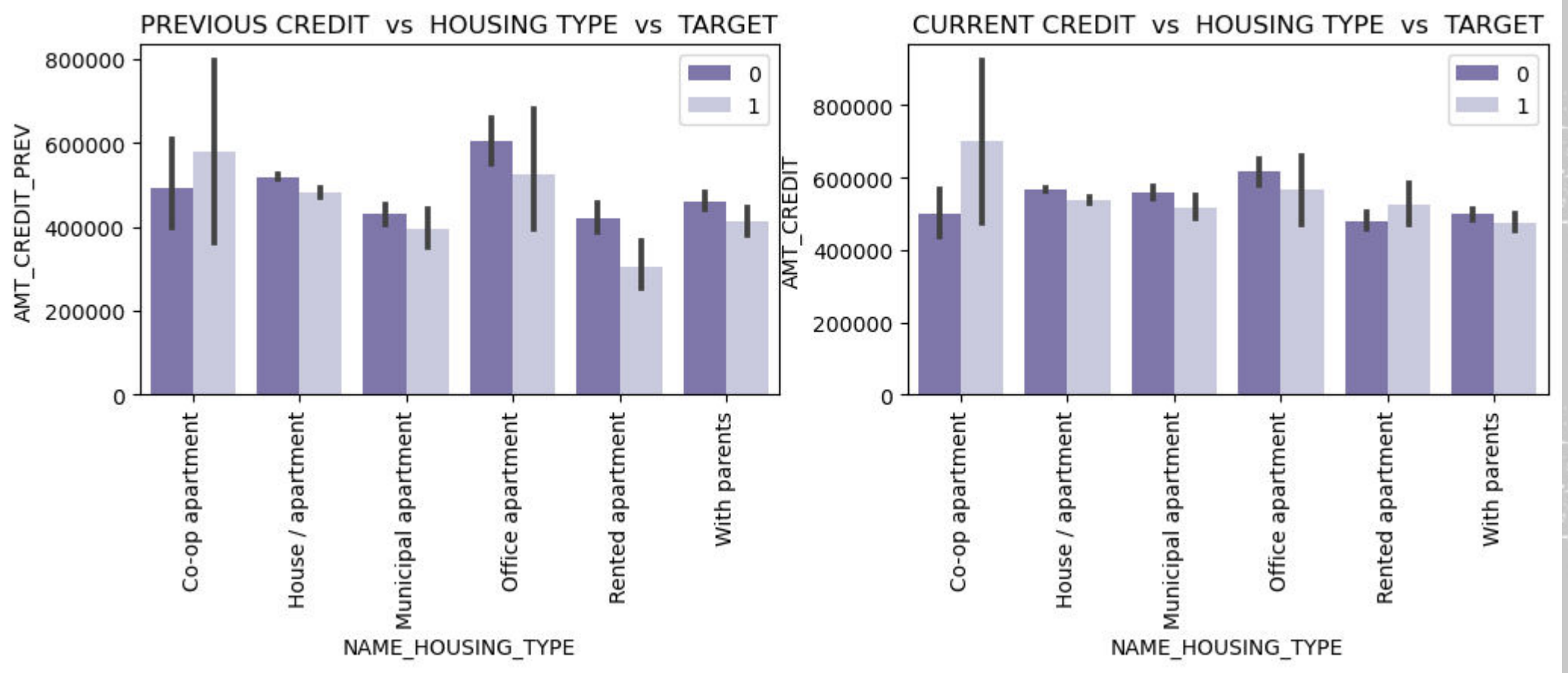
- Most of the clients whose **previous credit amount higher than 700k has canceled their contract**.
- **Approved** previous credit amount is **low** compared to refused previous credit amount.

- **Comparing previous credit , current credit and contract status in a single plot**



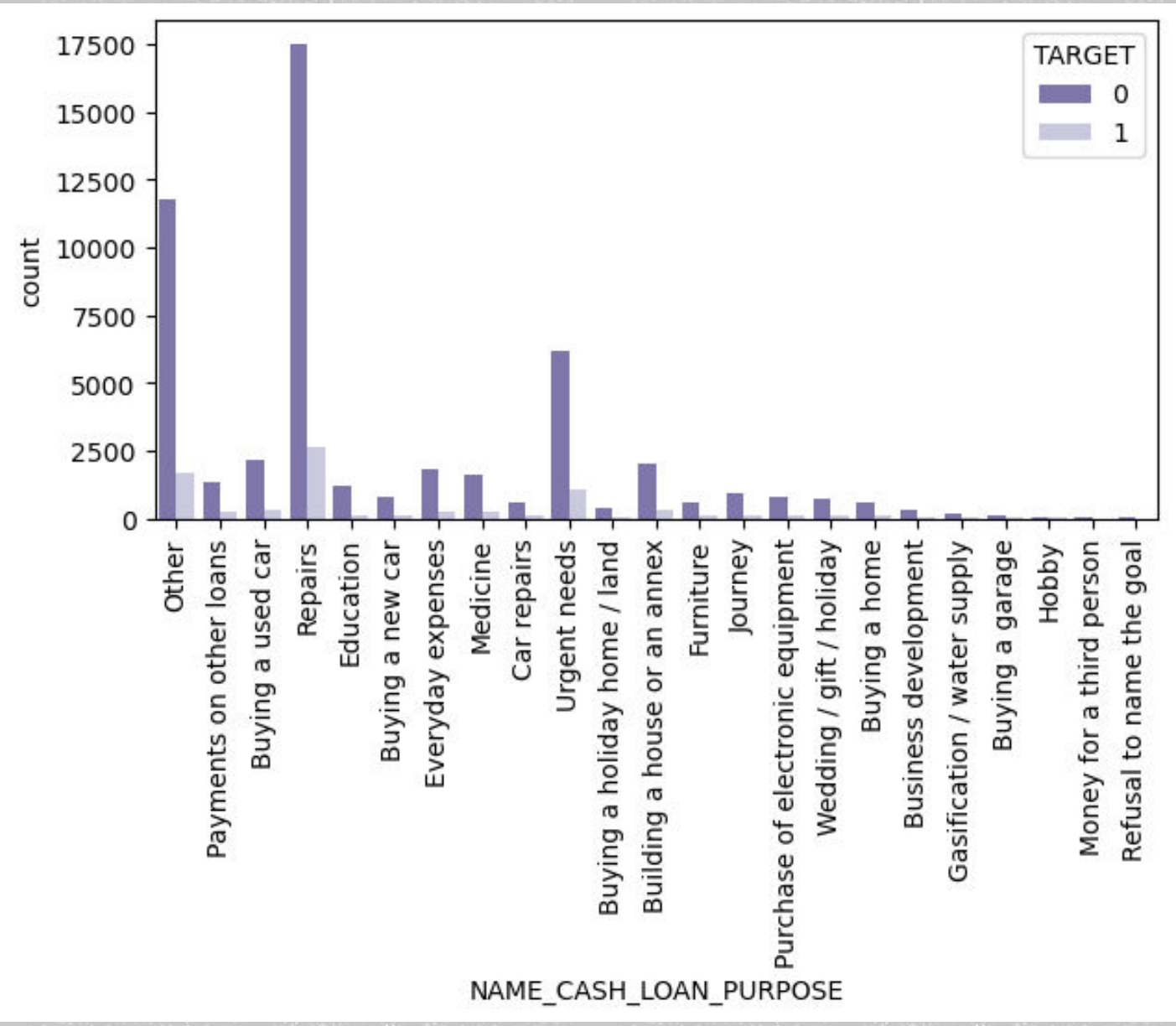
- **Unused offers** has **significantly increased compared to previous credit amount in range >600k.**
- **Canceled contracts have reduced a bit** compared to previous **credit range from ~800k to just above 600k.**
- **Approved contracts have significantly increased from previous credit range between 350k - 400k to >550k.**

Analysis on AMT_CREDIT_PREV , TARGET and NAME_HOUSING_TYPE



- There is a significant increase for credit range for customers in rented apartments and we are seeing an increase in difficulty for repaying loans in this category.
- Customers in office apartments are comparatively having less difficulty in repaying loans.
- Co-op apartment customers are having higher previous credit amount and also having difficulties in repaying loans.

Analysis on NAME_CASH_LOAN_PURPOSE categorized by TARGET



- Purpose of loan is maximum for repairs and comparing to the no. of successful repay on time, there is a low no. of payment difficulties.
- After repairing needs, customers take loans for urgent needs.

CONCLUSION

1. Credit amount is high for customers in co-op apartments while they are difficulties in paying loans on time. Customers in housing category with Parents are less likely to fail to repay.
2. Customers in rented apartments is showing increase in difficulty in paying loans back on time than not having difficulties.
3. Bank must find customers with parents as they are the ones with less number of defaulters comparatively.
4. Significant increase in unused offer/canceled in credit range over 600k.
5. Bank must look for customers with secondary special education as customers with this type of education repay loans on time.
6. Most of the customers who take loans are Laborers in occupation and repay loans on time. Nearly ~10% of these people find it difficult to repay on time.
7. Most of the customers taking loans are in middle age group (35 to 60) and repay on time. Only ~7% of this category find it difficult to repay loans.
8. Most of the customers taking loans are married and repay loans on time. Only ~7.5% of this category find it difficult to repay loans.
9. More than 50% of the total customers are in "Working" category and in that only ~9.5% are finding it difficult to repay loans.
10. Purpose of loan is maximum for repairs and comparing to the number of successful repay on time, there is a low number of payment difficulties.