# CREDIT EDA CASE STUDY

Group No 12

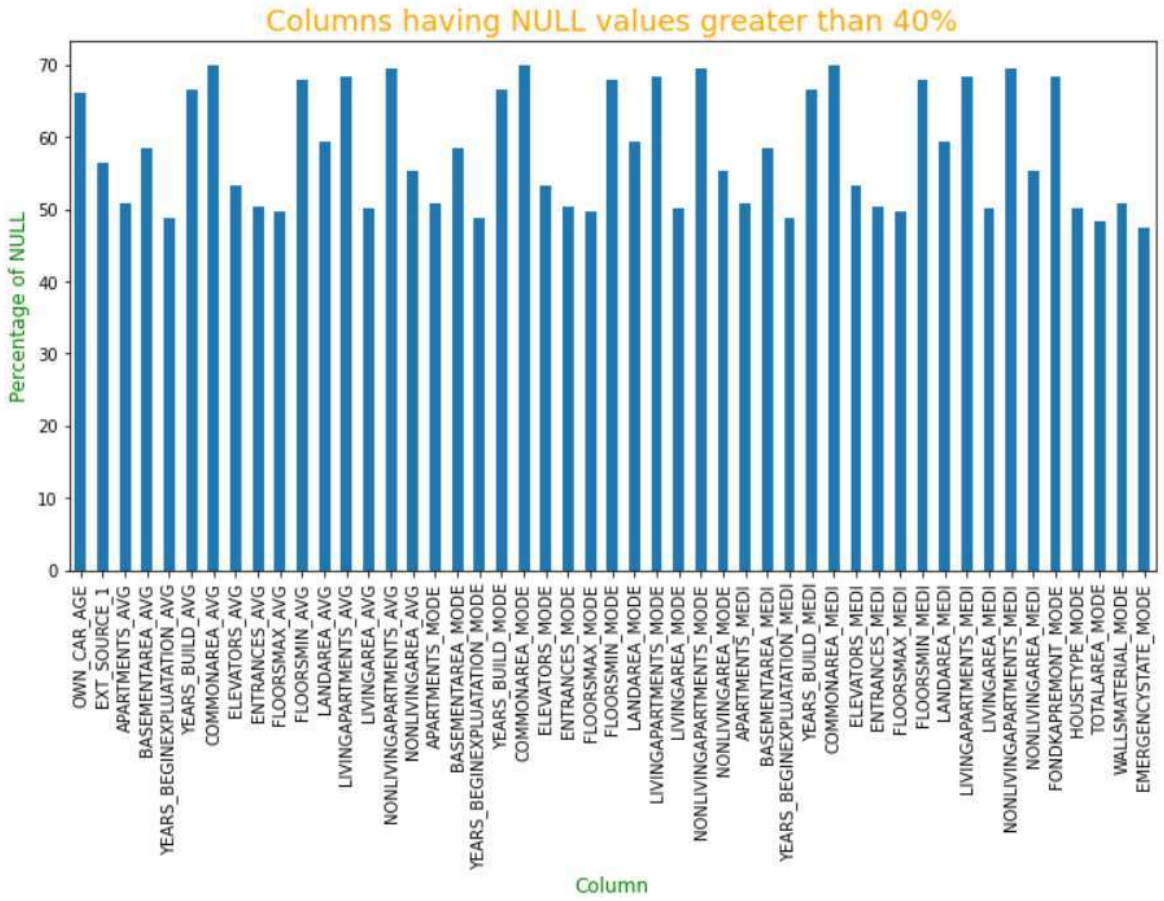
Vaishnav Diliprao Jadhav
Varun Vishweshwar Chitale
Vishal Rajendra Khetmalis
Vishvnath Hambire
Yogesh Shriramji Nakhate

# Problem Statement

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.

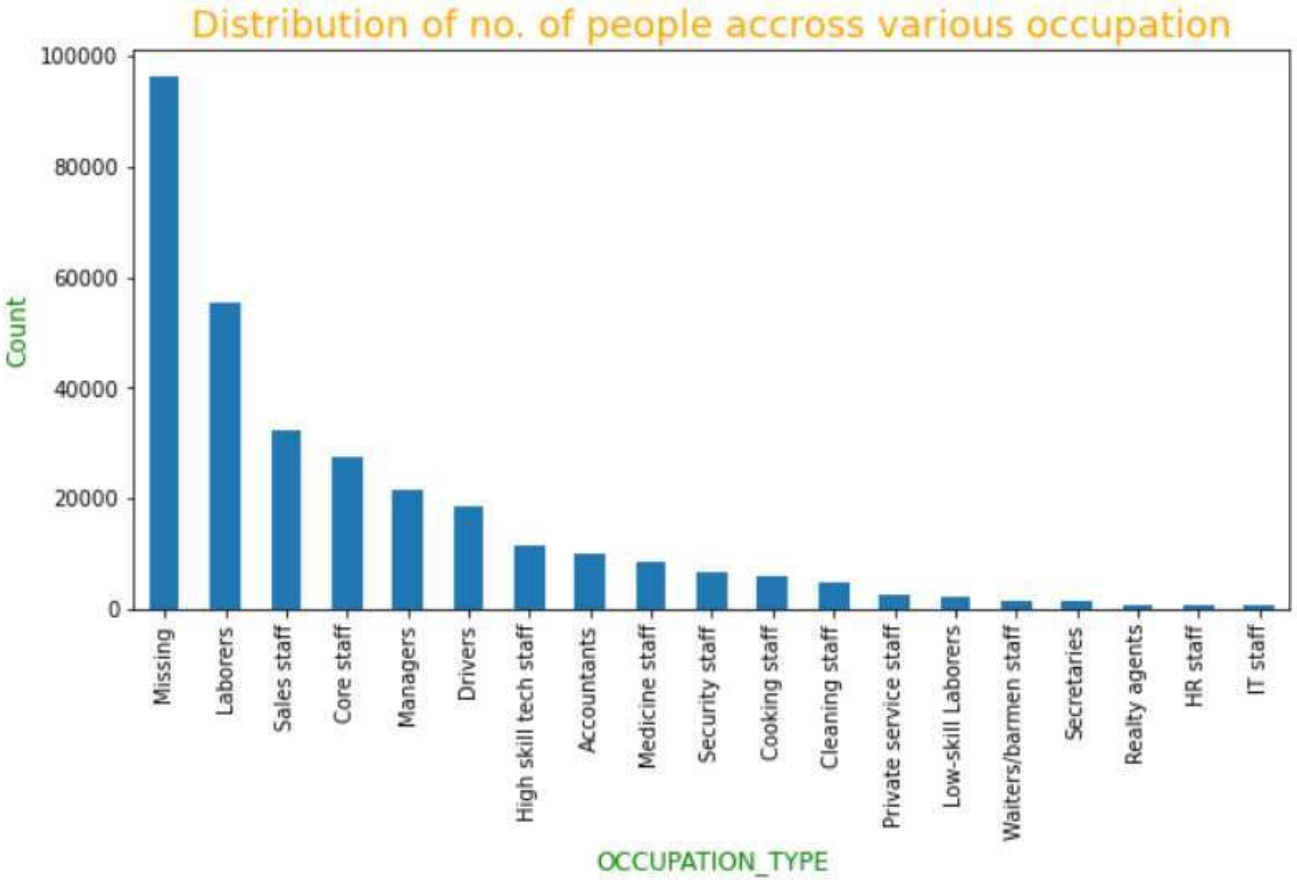
# Identifying and Treating NULL values




The graph on the left side depicts:

- The columns which has **null percentage** greater than **40%** .
- In total there are **49 columns** and we will be dropping these columns to focus more one important columns.
- Some of the columns which are removing are COMMONAREA_AVG,YEARS_BUILD_AVG etc.

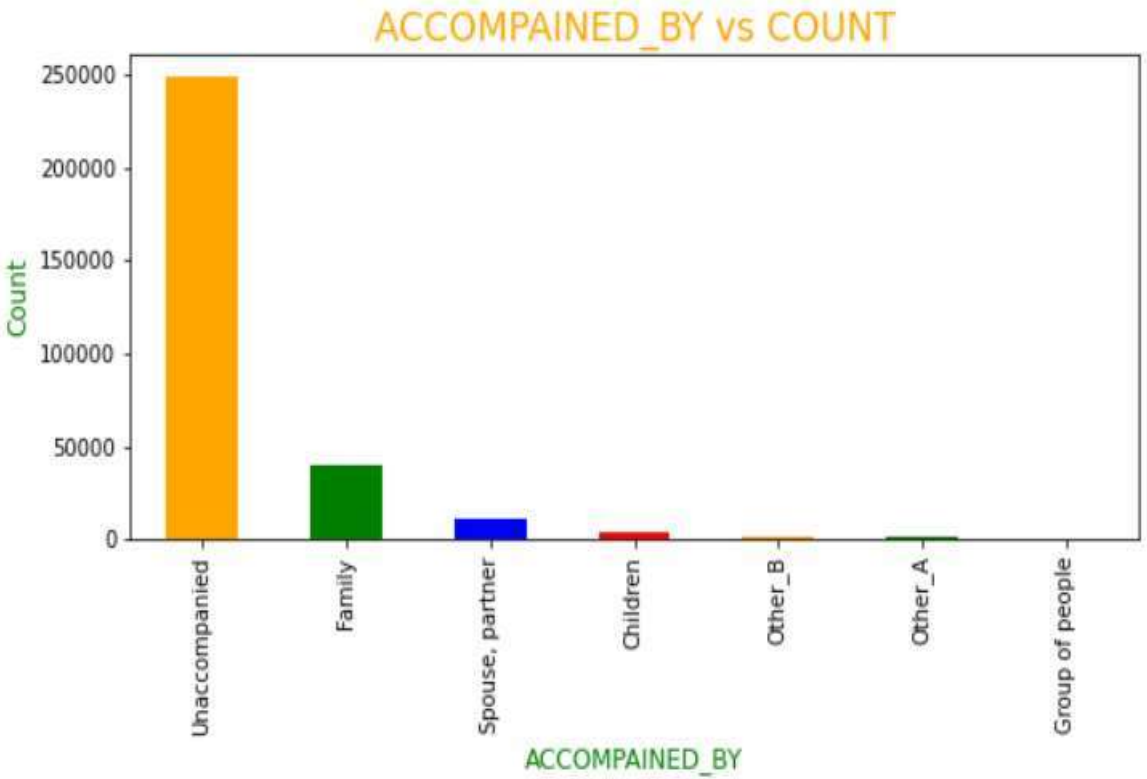
# Treating NULL values of <u>OCCUPATION_TYPE</u>




After analysing OCCUPATION_TYPE we found that

- The column has **31% NULL** values.
- As OCCUPATION_TYPE can be an important parameter in our analysis, it wouldn't be wise to drop it.
- Imputing it with some other value such as mode can also make our inferences biased.
- We have created a separate category "**Missing**" for it.

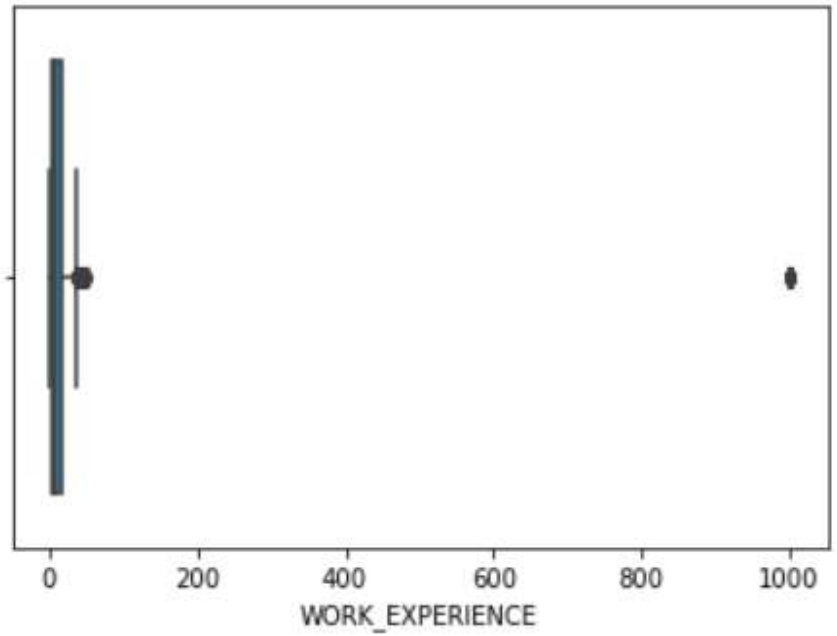
# Treating NULL values of NAME_TYPE_SUITE




After analysing NAME_TYPE_SUITE we found that

- There are some missing values in NAME_TYPE_SUITE(**0.42%**)
- From the graph on left side, we can see that most of the time applicant was Unaccompanied.
- We can use **mode** to impute the missing values as the missing value count is very less, it won't have larger impact.

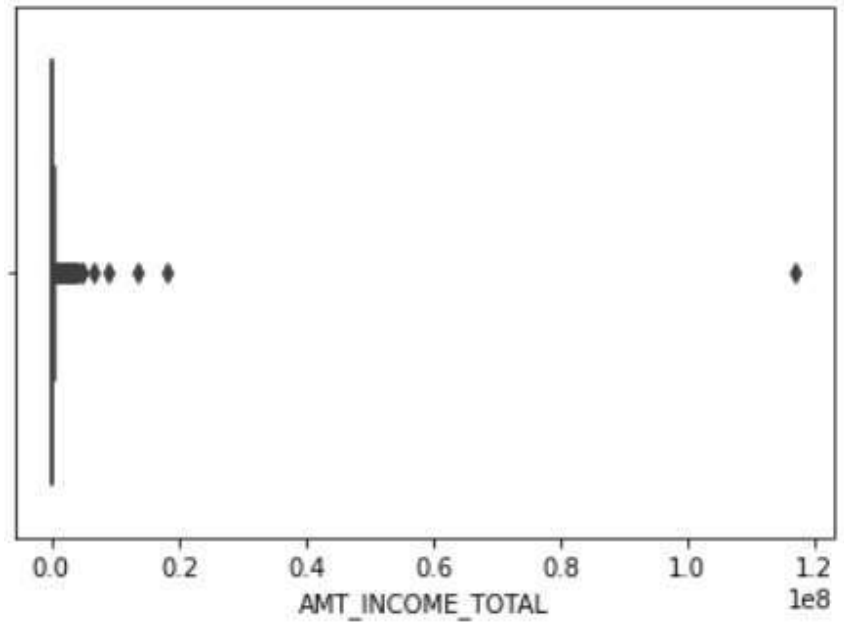
# Checking Outliers in WORK_EXPERIENCE




Boxplot on the left side depicts as below:

- We have plotted box plot for WORK_EXPERIENCE(DAYS_EMPLOYED converted to years)
- We can see a set of value which is higher than **999**.
- It is not a possible scenario, hence these values are outliers and we will be **excluding** it from our analysis.

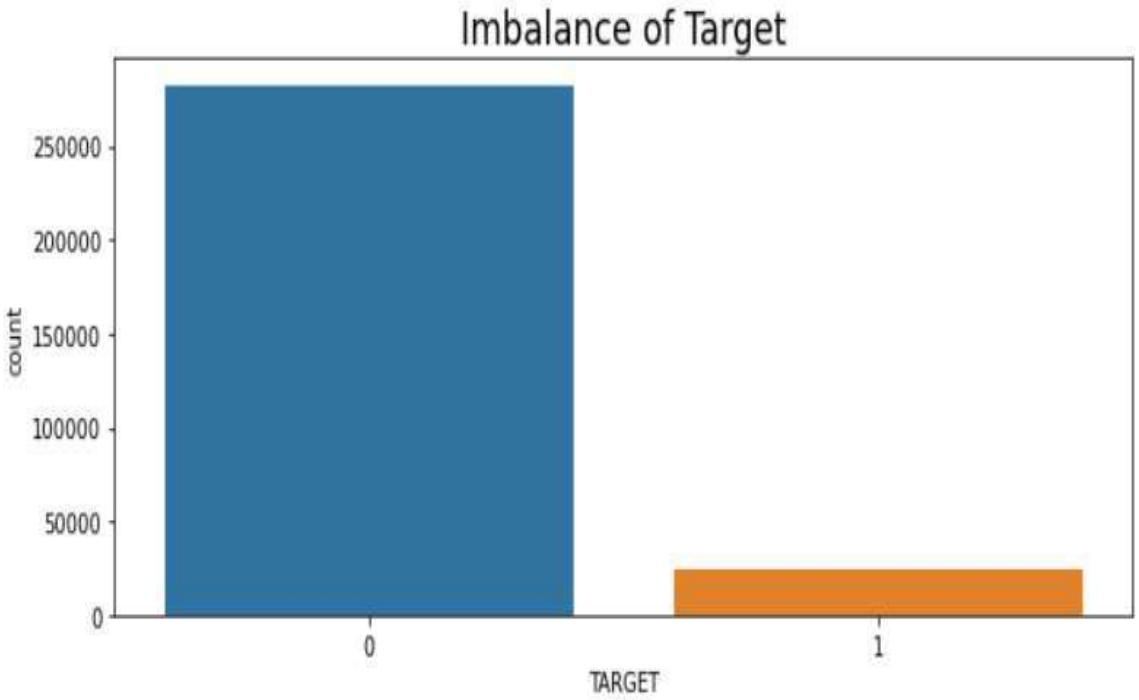
# Checking Outliers in AMT_INCOME_TOTAL




Boxplot on the left side depicts as below:

- We have plotted box plot for AMT_INCOME_TOTAL.
- We can see a outlier, after inspecting the outlier we observed that it is the **maximum value** in AMT_INCOME_TOTAL.
- The AMT_INCOME_TOTAL belongs to **labourer** and his **target** variable is also **1**.
- This is not a possible scenario. Hence we will be **dropping** this row.

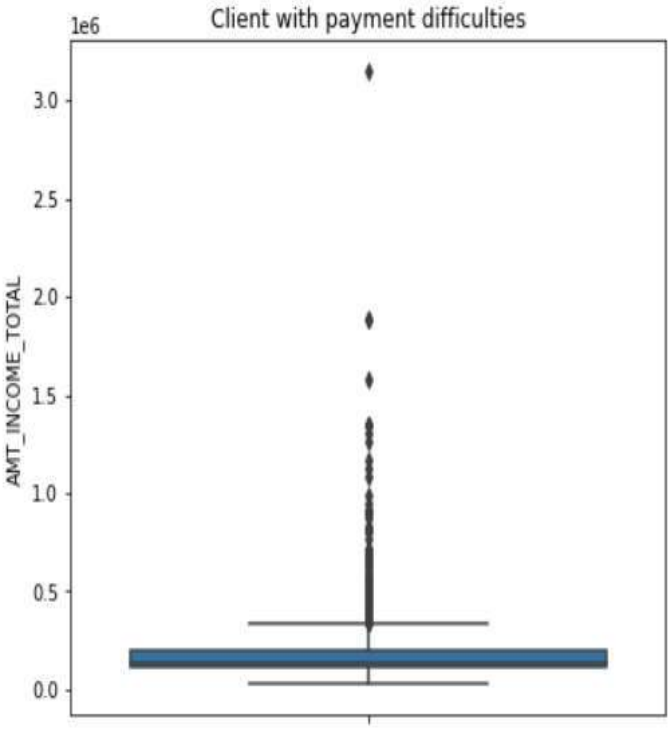
# Target Imbalance

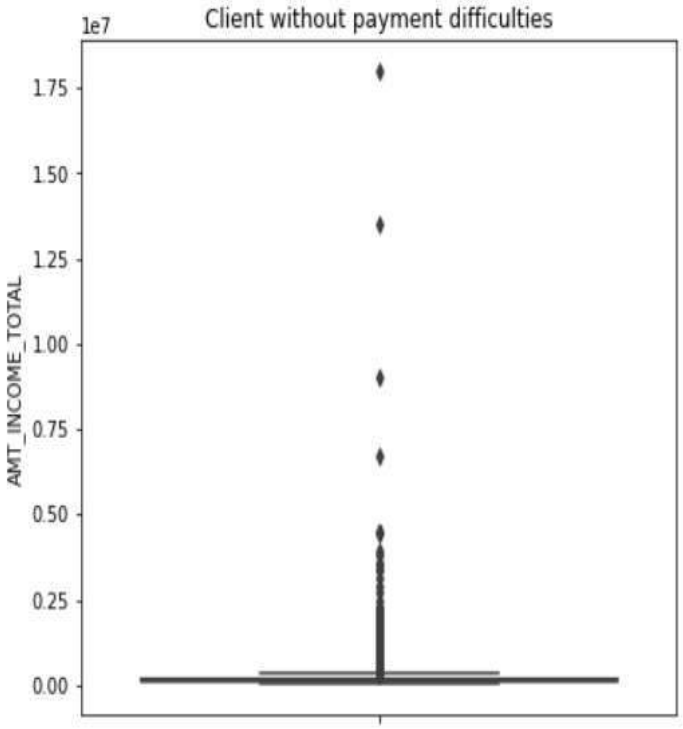



The bar chart to the left depicts as below:

- The Bar chart describes about the imbalance of the Target variable.
  0 - Applicant's with no payment difficulties
  1 – Applicant's with payment difficulties
- Target variable **0 hold's approx. 91%**
- Target variable **1 hold's approx. 8%**
- We can clearly see data imbalance, we will **divide dataset** into two for further analysis.

# Univariate Analysis **AMT_INCOME_TOTAL**

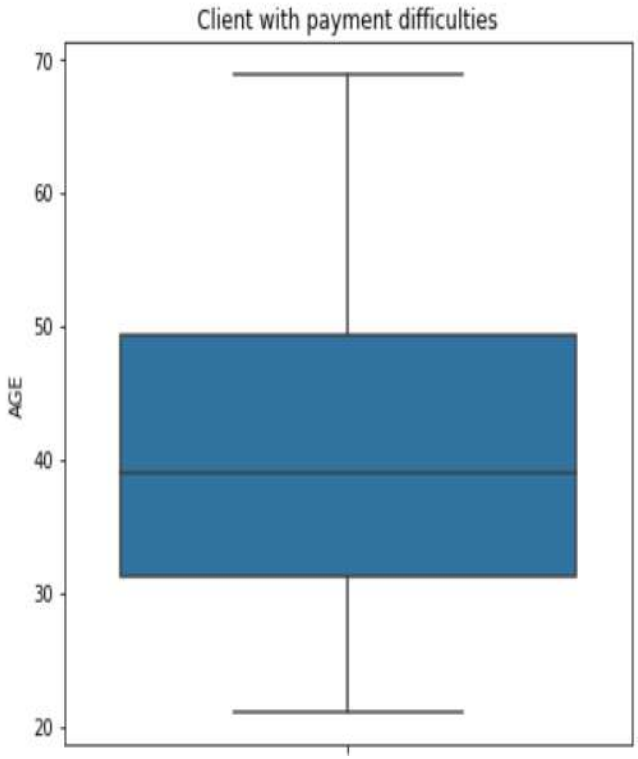


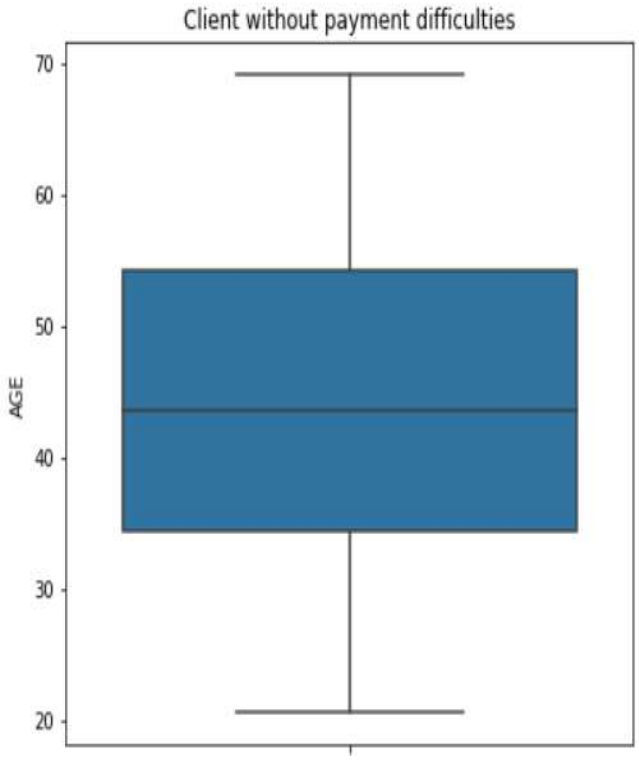




The box plot one left depicts

- It represents univariate analysis of **AMT_INCOME_TOTAL**.
- First plot represents the customer with payment difficulties
- Second plot represents customer without payment difficulties
- We can infer that total income is higher for client without payment difficulties as compared to client with payment difficulties

# Univariate Analysis **AGE**

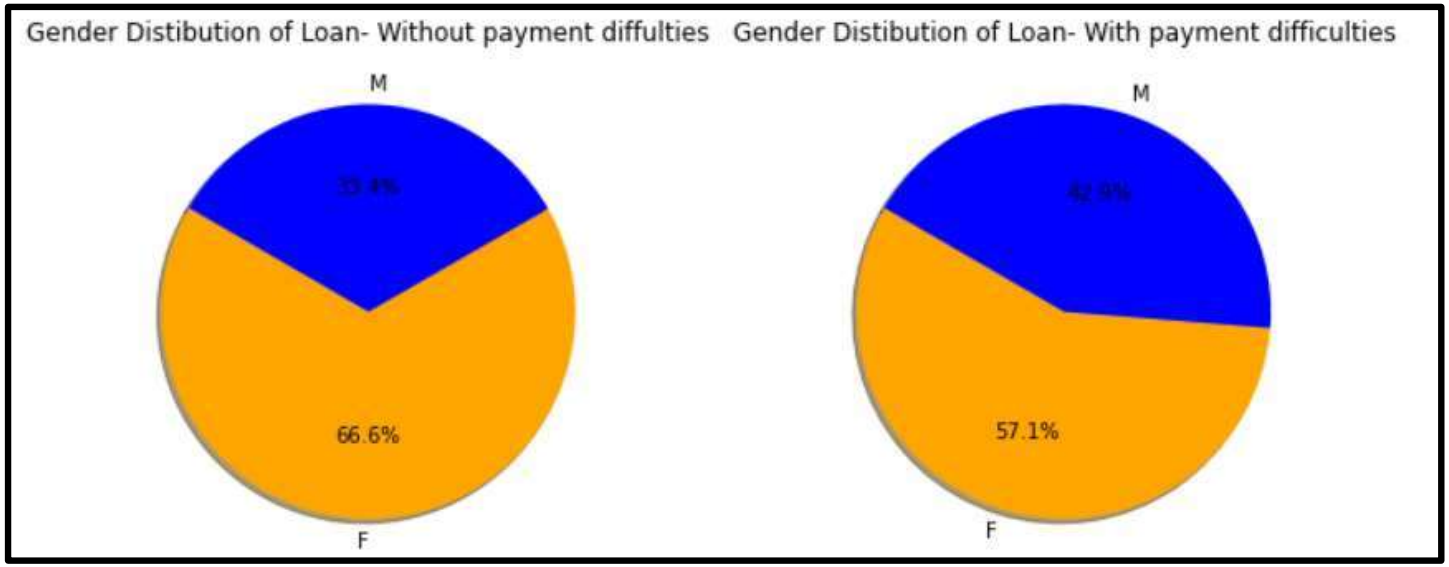



The box plot one left depicts

- It represents univariate analysis of **AGE**.
- First plot represents the age of customer with payment difficulties
- Second plot represents age of customer without payment difficulties.
- By observing the boxplot, we can infer that Client with payment difficulties are in range of 31-49.
- client without payment difficulties are in range 34-54

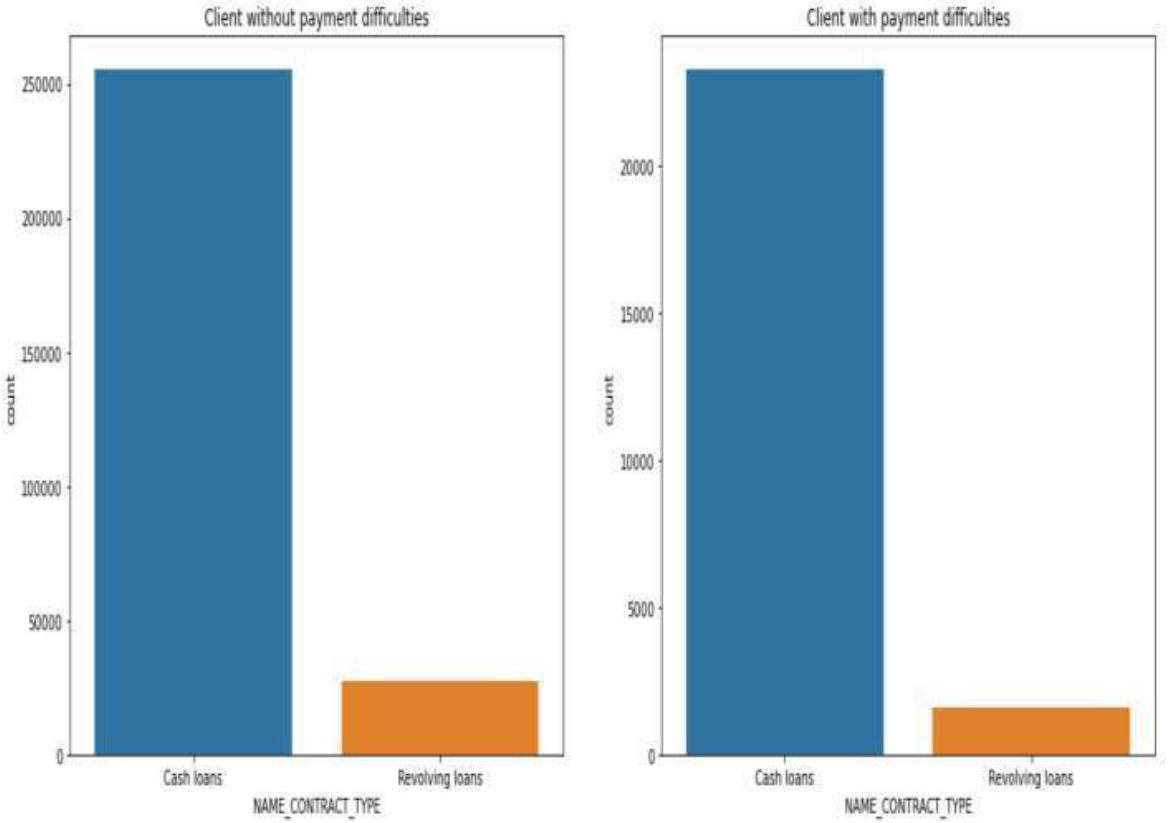
# Univariate Analysis of **categorical variable**




The above pie chart depicts:

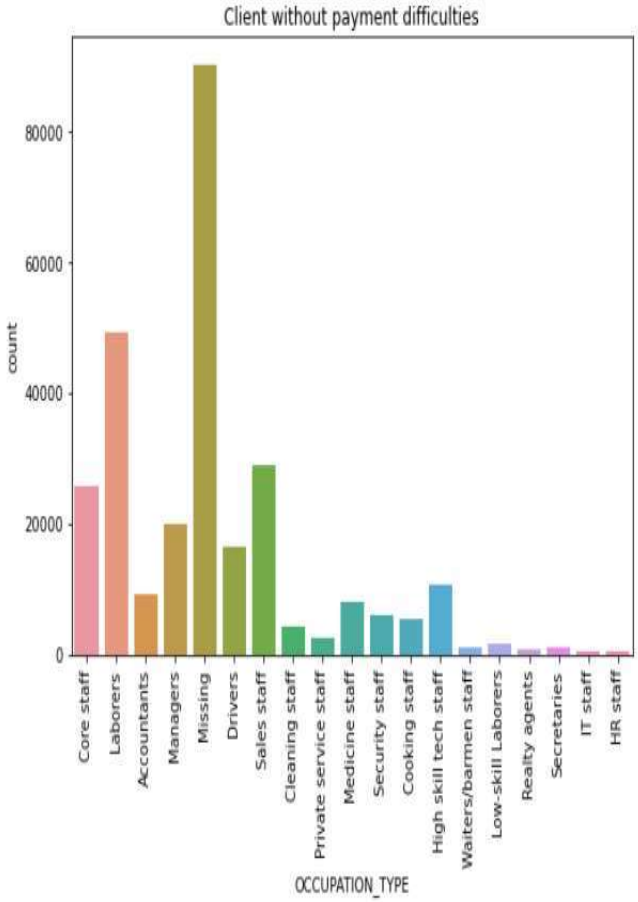
- It shows the distribution of male and female population, with payment difficulties and without payment difficulties
- Let's consider the second pie chart with payment difficulties. We can see that **Female** account for **57.1%** and **Male** account for **42.9%**
- Here we observe that **Female** have **more payment difficulties** as compared to male

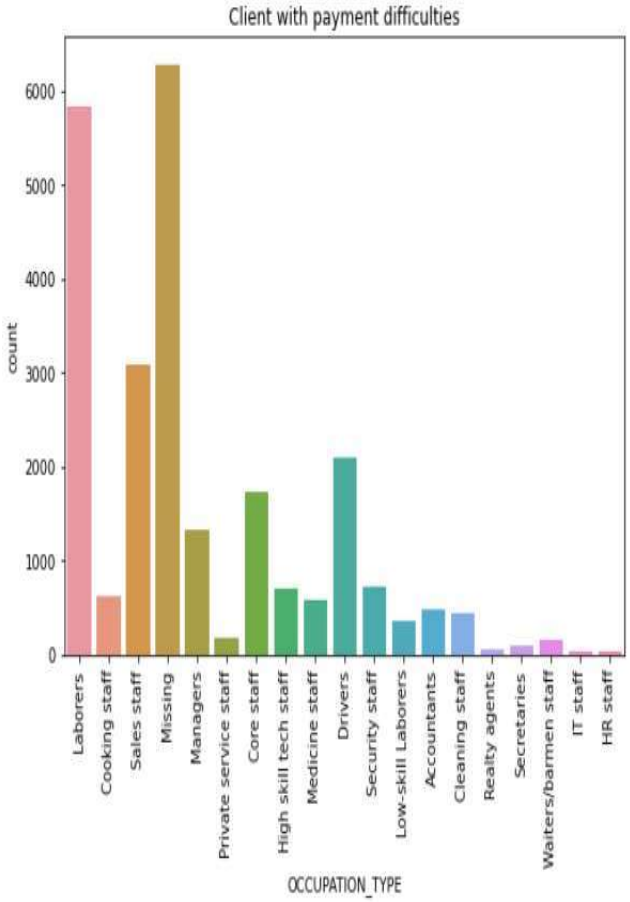
# Univariate Analysis on **NAME_CONTRACT_TYPE**




- The figure to the left depicts as below:
  - The count plot represents the Univariate analysis on the **NAME_CONTRACT_TYPE** column.
  - First plot represents that customers without any payment difficulties majorly apply for Cash loans and followed by Revolving loans.
  - Whereas the second plot represents that the customers with payment difficulties also majorly belong to Cash loans followed by Revolving loans.
  - There **isn't much difference** in both the plot's

# Univariate Analysis on **OCCUPATION_TYPE**

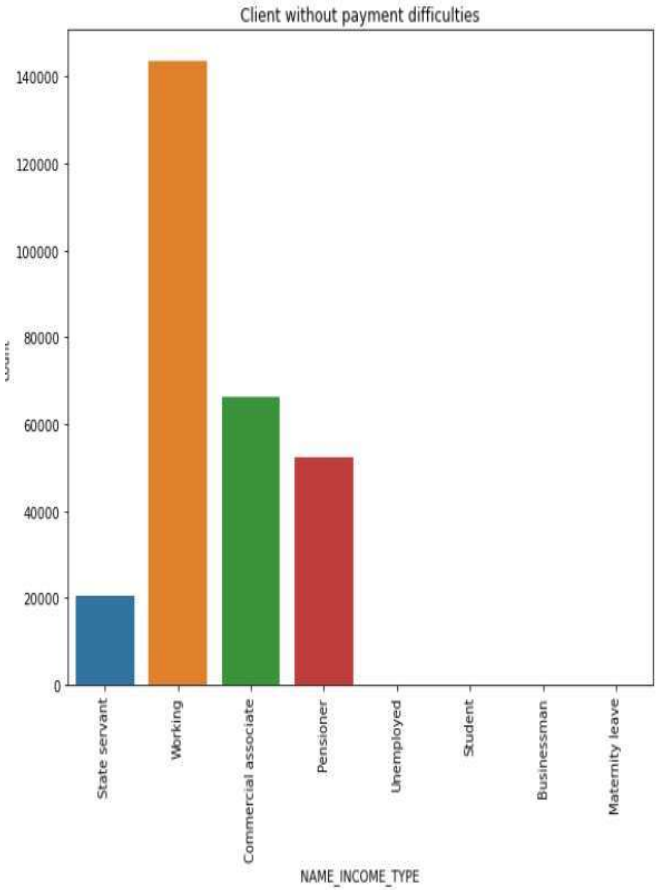


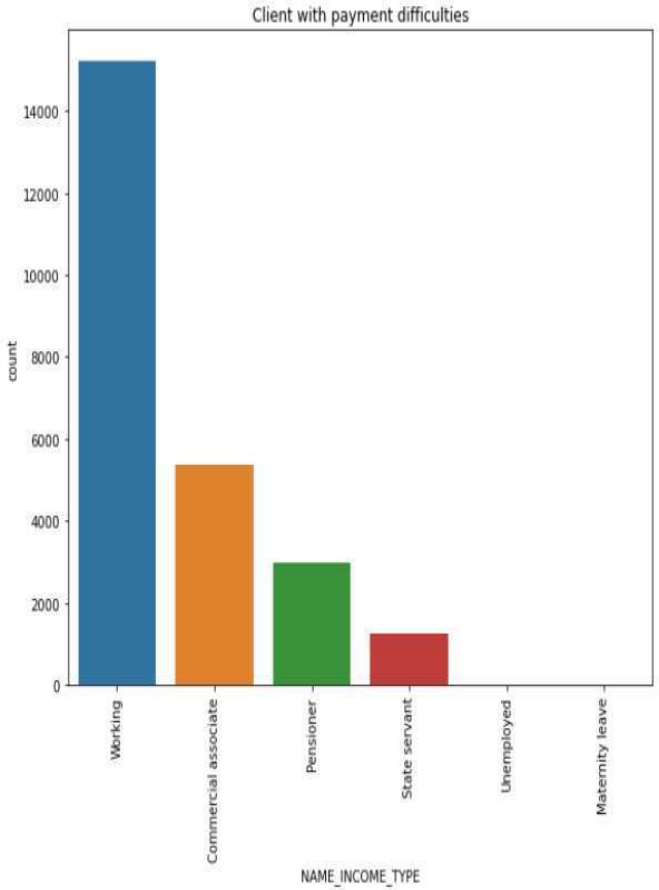




The count plot one left depicts

- The plot represents univariate analysis on **OCCUPATION_TYPE**
- The first plot represents distribution count of occupation type for applicants without payment difficulties.
- The second plot represents distribution count of occupation type for applicants with payment difficult.
- In both the plot's we can observe that **labourer's** have the **highest count** followed by Sales staff.

# Univariate Analysis on **NAME_INCOME_TYPE**

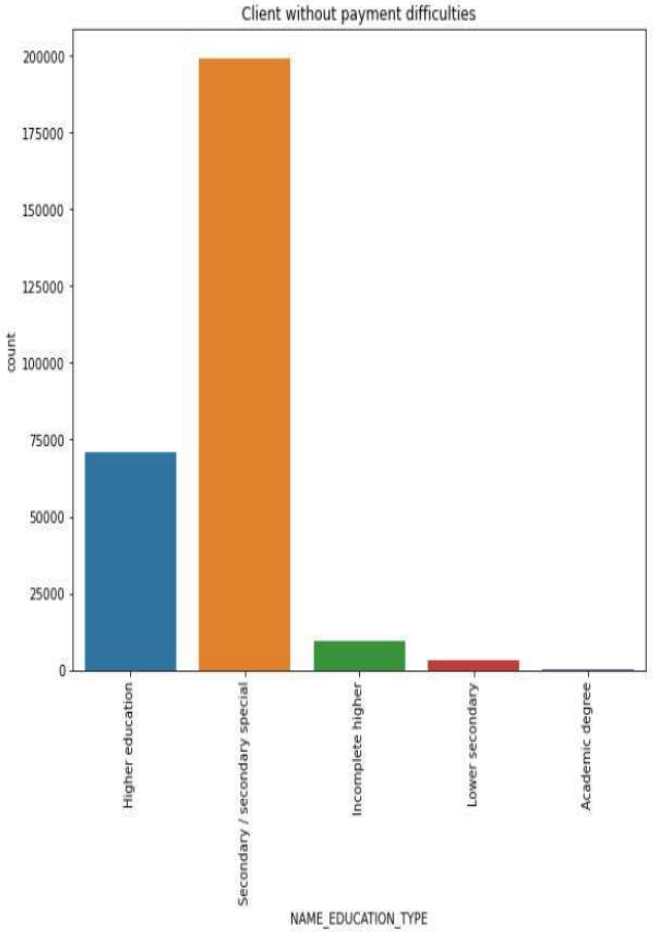


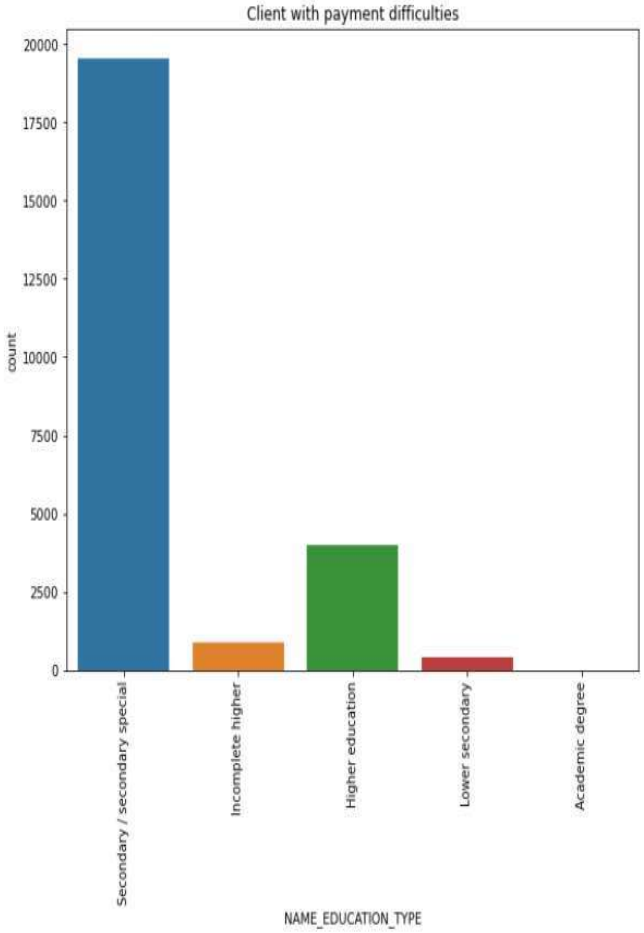




The count plot one left depicts

- The plot represents univariate analysis on **NAME_INCOME_TYPE**
- The first plot represents distribution count of Income type of applicants without payment difficulties.
- The second plot represents distribution count of Income type of applicants with payment difficult.
- From the plot we can infer that **Pensioner** and **Govt Employees** have **better on-time payments**.

# Univariate Analysis on **NAME_EDUCATION_TYPE**

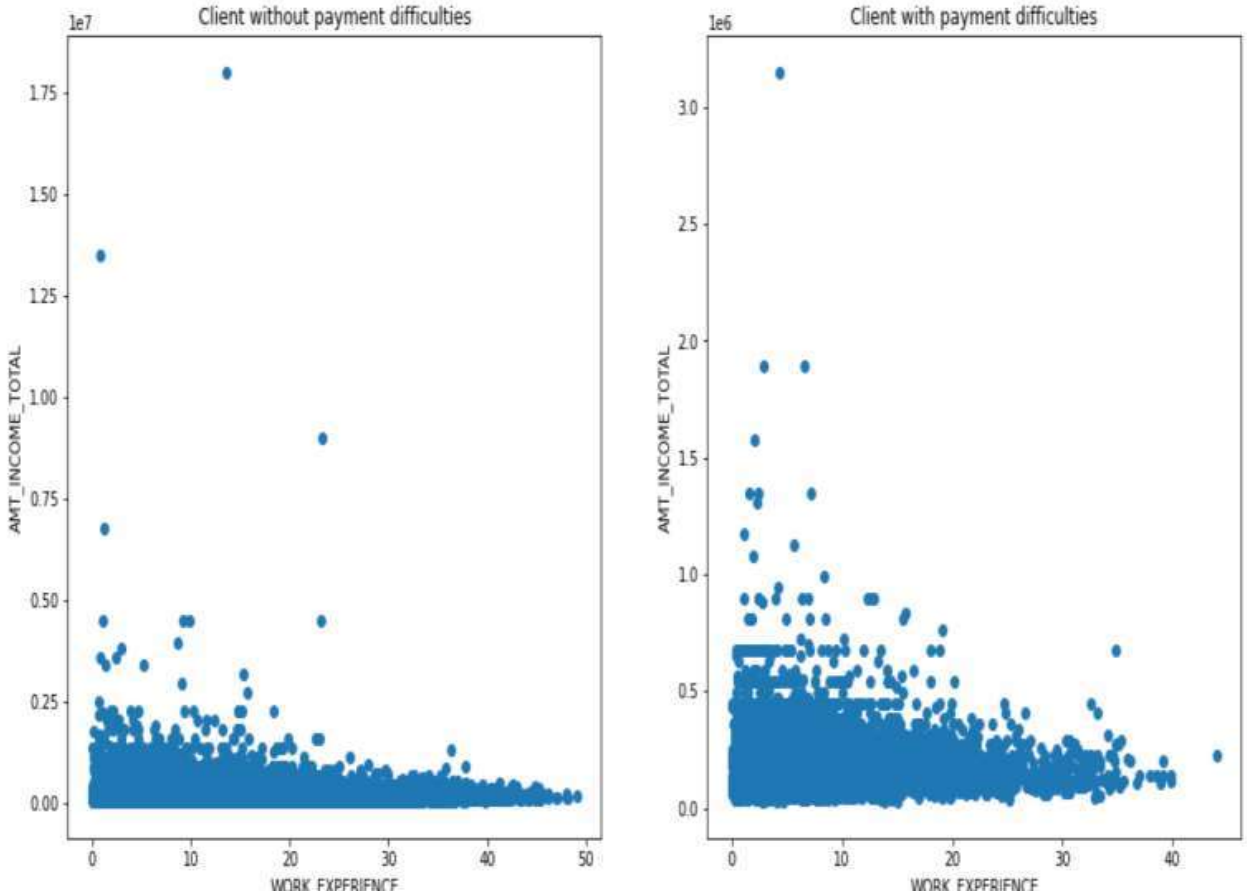



The count plot one left depicts

- The plot represents univariate analysis on **NAME_EDUCATION_TYPE.**
- The first plot represents distribution count of Education type of applicants without payment difficulties.
- The second plot represents distribution count of Education type of applicants with payment difficult.
- From the plot we can infer that count of customers who have completed **secondary/secondary special** is highest for both target variables.

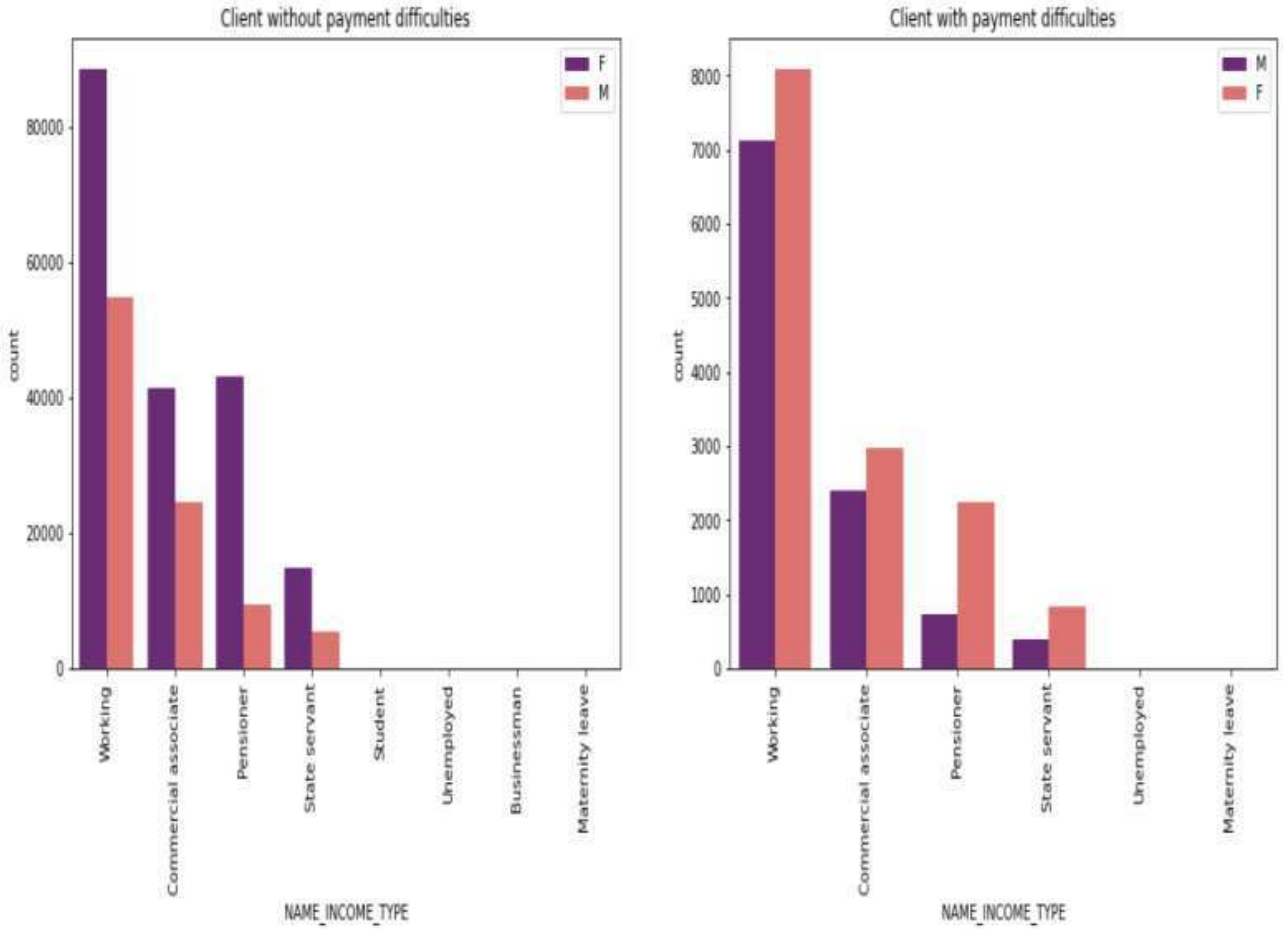
# Bivariate Analysis on **YEARS_EMPLOYED** and **AMT_INCOME_TOTAL**




The scatterplot on the left depicts as below:

- The plot represents bivariate analysis between **AMT_INCOME_TOTAL** and **WORK_EXPERIENCE**.
- In first plot we can observe that as the WORK_EXPERIENCE increases there is decrease in AMT_INCOME_TOTAL
- In the second we can observe similar trend i.e with increase in WORK_EXPERIENCE, AMT_INCOME_TOTAL is decreasing.

# Bivariate Analysis of **GENDER** and **NAME_INCOME_TYPE**



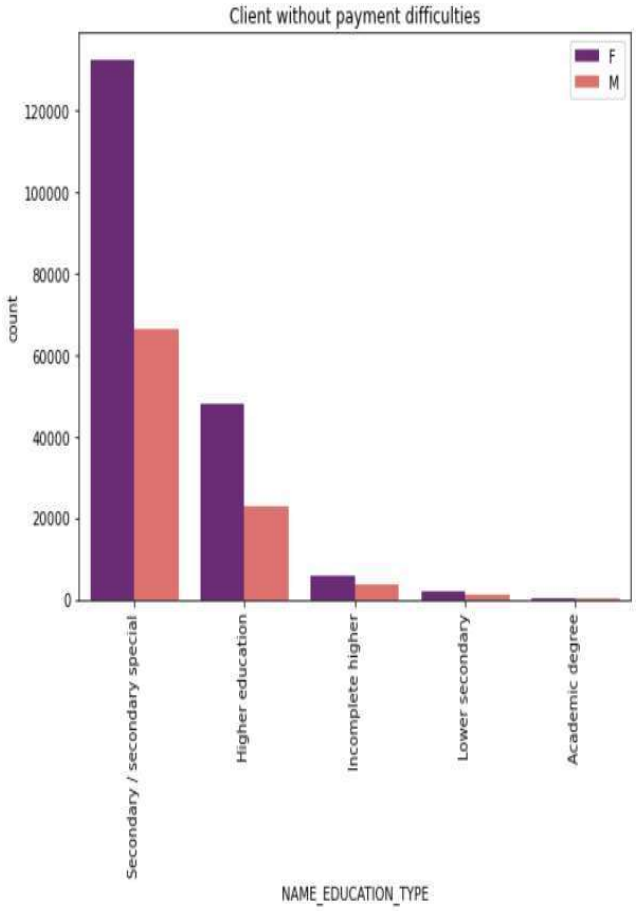


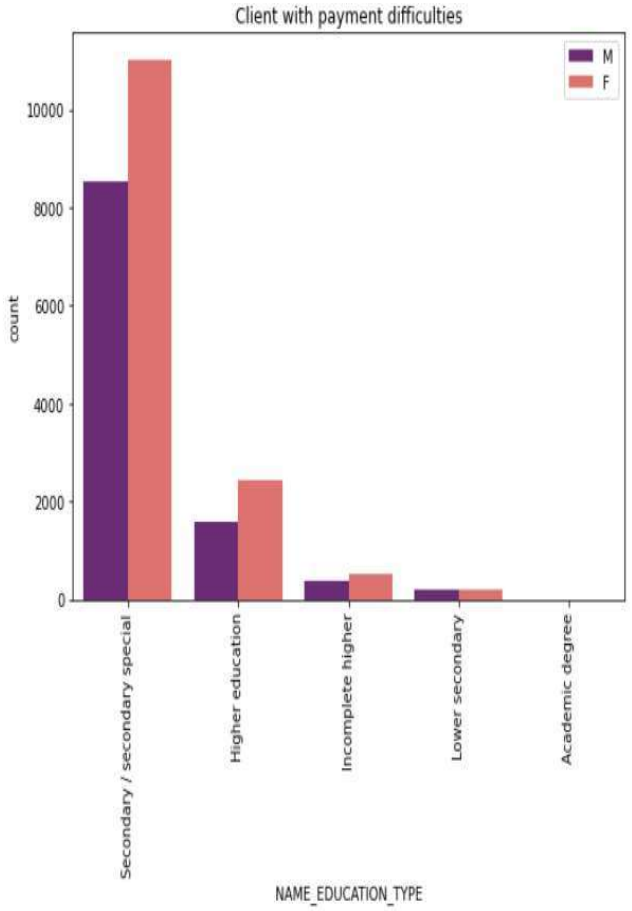


The scatterplot on the left depicts as below:

- The plot represents bivariate analysis between **NAME_INCOME_TYPE** and **GENDER**.
- Female applicant's have more difficulties in payment as compared to male applicant’s.
- Applicant's who are businessman and student's pay their loan on time although there count is low.

# Bivariate Analysis of **GENDER** and **NAME_EDUCATION_TYPE**

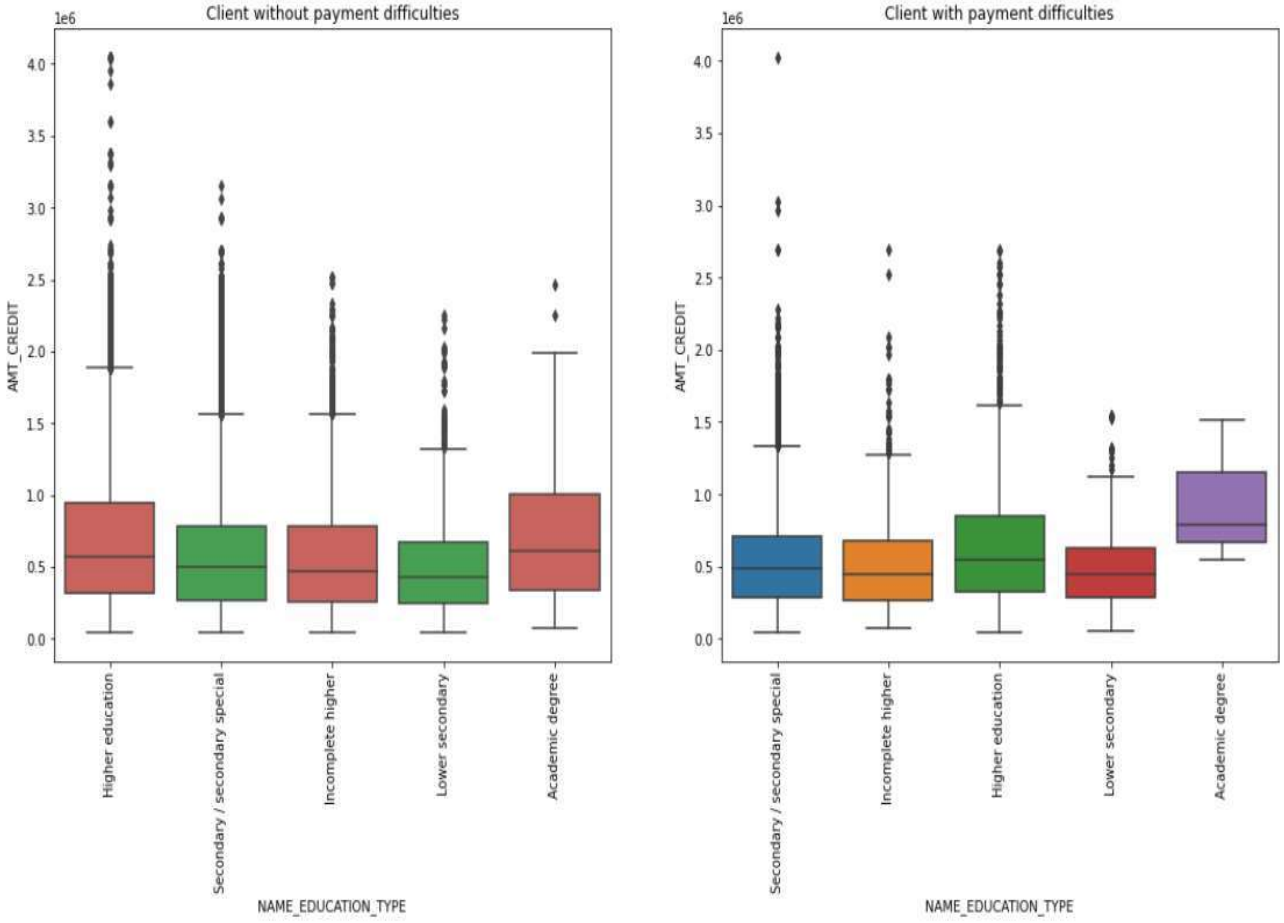



The scatterplot on the left depicts as below:

- The plot represents bivariate analysis between **NAME_EDUCATION_TYPE** and **GENDER**.
- Applicant's who have higher education have less difficulty in paying loan as compared to Secondary/secondary special across both the gender’s.

# Bivariate Analysis on **AMT_CREDIT** and **NAME_EDUCATION_TYPE**



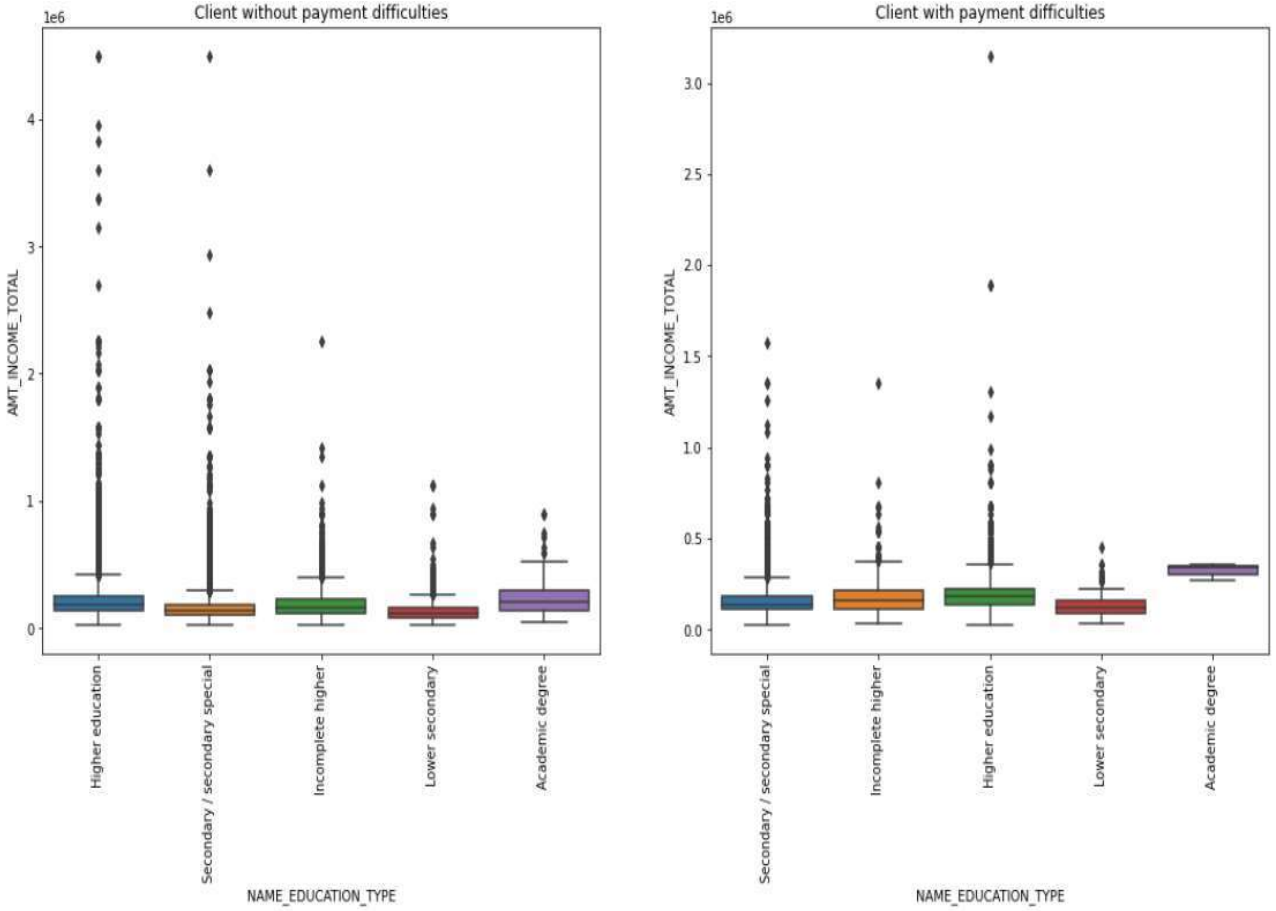




The box plot on the left depicts as below:

- The box plot represents the Bivariate analysis on the **AMT_CREDIT** column and **NAME_EDUCATION_TYPE** column.
- We can observe in first plot that there are more outliers for Higher education and then it decreases for other education type.
- In case of customer's with payment difficulties, outliers are more for secondary/secondary special and then it decreases for other education_tpes

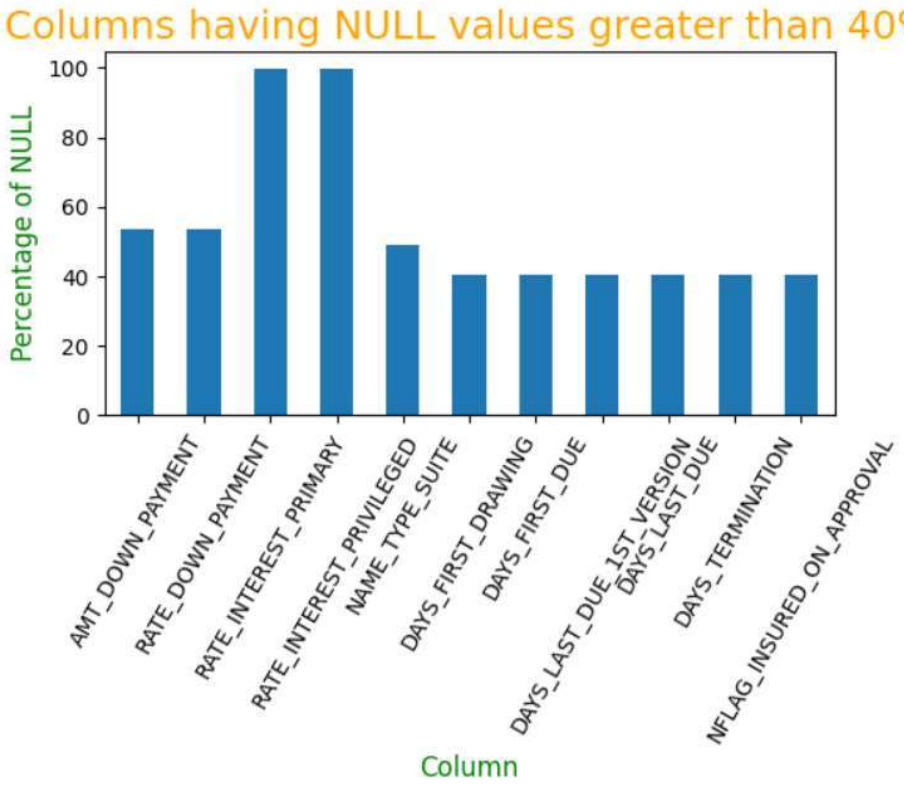
# Bivariate Analysis on **AMT_INCOME_TOTAL** and **NAME_EDUCATION_TYPE**




The figure to the left depicts as below:

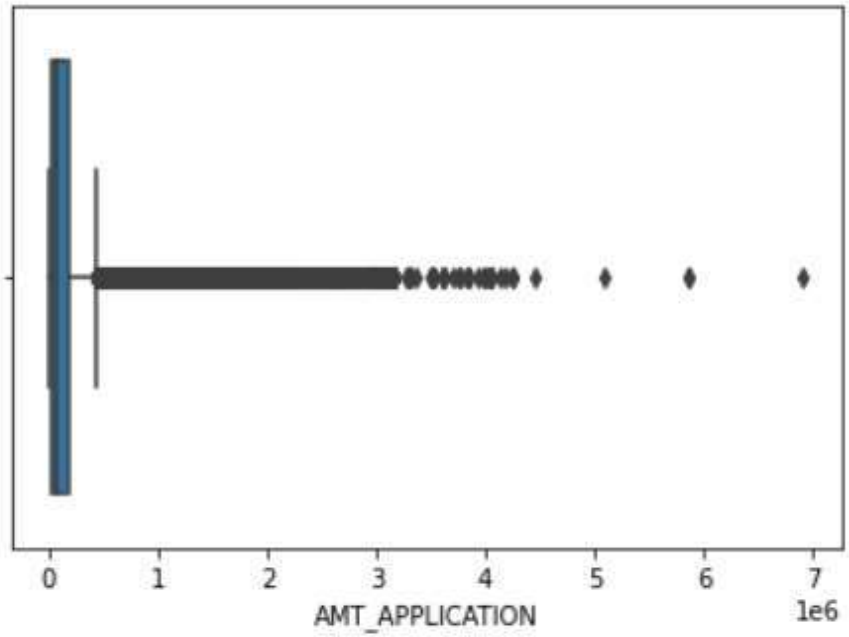
- The box plot represents the Bivariate analysis on the **AMT_INCOME_TOTAL** column and **NAME_EDUCATION_TYPE** column.
- The first plot depicts that customer without payment difficulties have higher outliers for Higher education
- The second plot depicts that customer with payment difficulties have lesser outliers.

# Working on **Previous_app** DataFrame




- **Previous_app** DataFrame consist of the id of the customers and their previous loan application related details like **CREDIT_AMT**, **APPLICATION_AMT** and **NAME_CONTRACT_STATUS.**
- The graph on left side showing the columns with more than 40% null values.
- There are approximately 11 columns which have null values more than 40% null values.
- To better analyse the DataFrame we will remove all these columns.
- Also we will remove some unused columns such as **HOUR_APPR_PROCESS_START** and **NAME_CASH_LOAN_PURPOSE** as it will not come in use in further analysis**.**

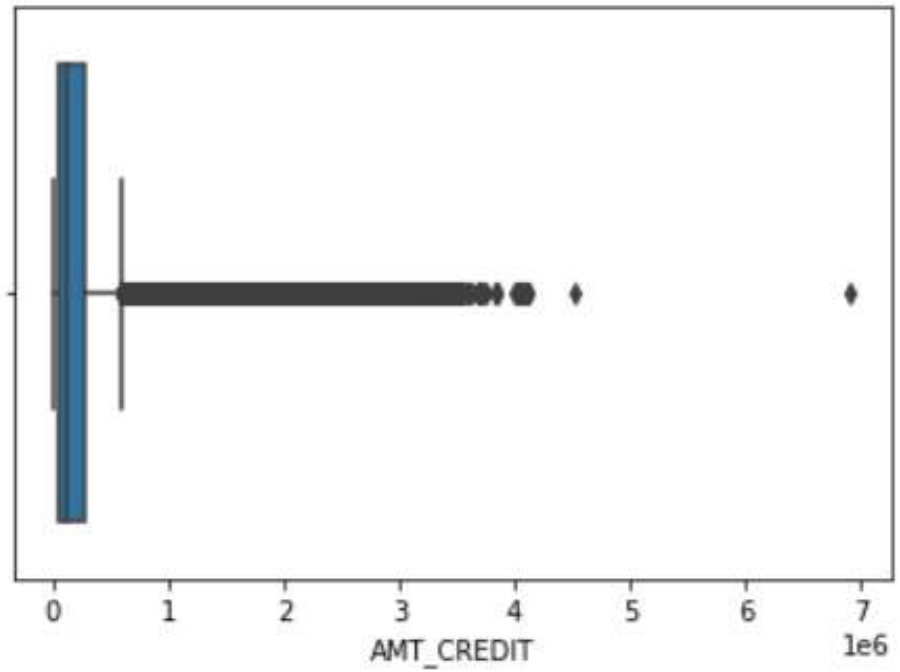
# Outlier handling in **AMT_APPLICATION**




- After plotting the boxplot of **AMT_APPLICATION** columns we can clearly see that there are some application amounts which are so much high but there count is not that much as but if we let them remain in the DataFrame that will affect our analysis.

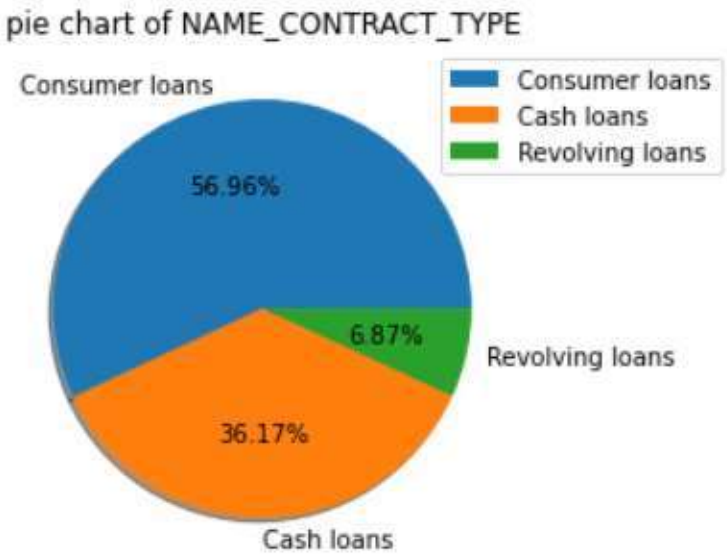
- For this purpose we will remove them from the DataFrame by taking only **99$^{th}$ percentile** values from the Columns.

# Outlier handling in **AMT_CREDIT**




- By analysing **AMT_CREDIT** column using the boxplot we can see that it also have some outliers which can affect our analysis.
- To remove them we will follow the same process and take the top **99$^{th}$ percentile** values and will remove the remaining values.
- After handling the outliers and cleaning the null values now we can go ahead with the analysis process.

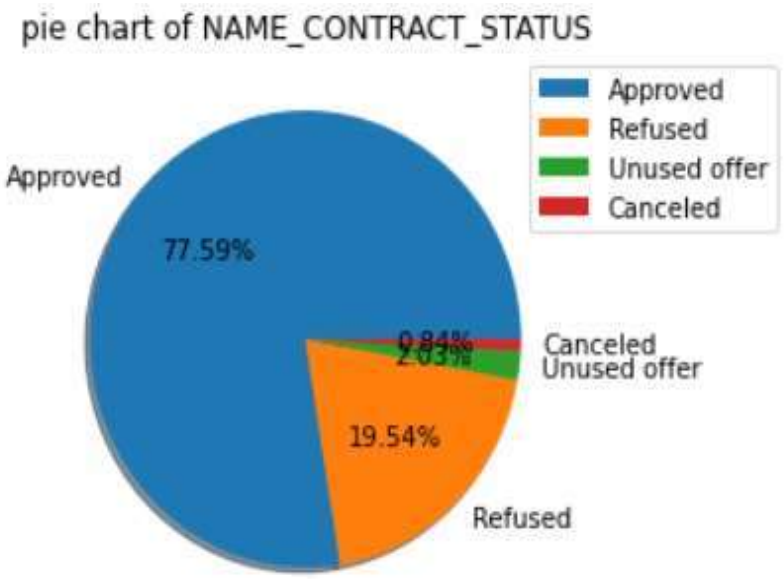
# Univariate Analysis on **NAME_CONTRACT_TYPE**




**INFERENCES:**

- Maximum application received by the banks are (approx. 57%) are for the consumer loans.
- Cash Loans percentage is 37% from the total loan applications.
- only 7% people demands for Revolving Loans which is the least percentage

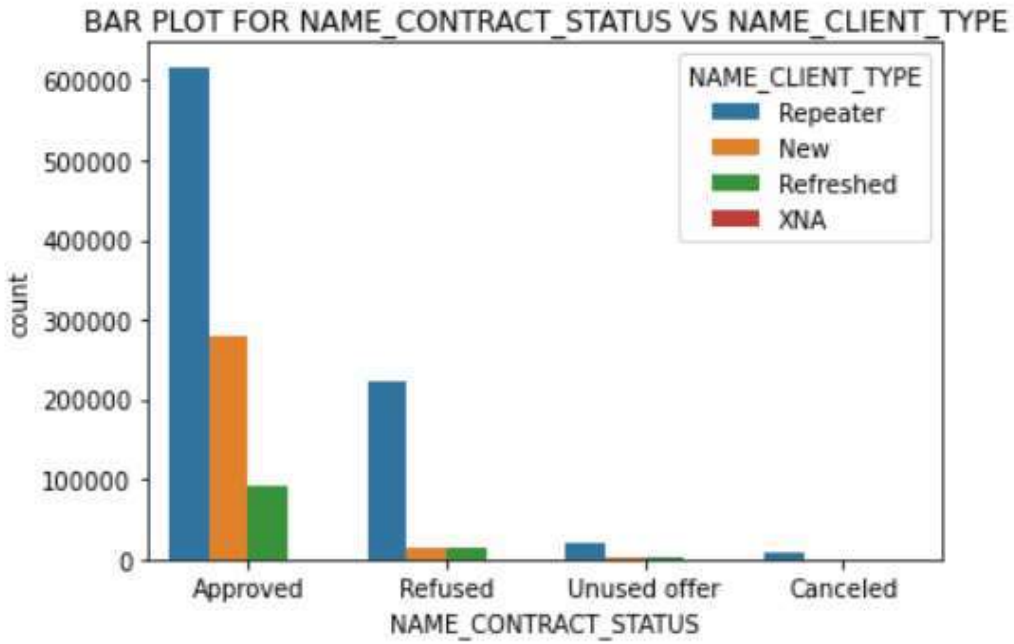
# Univariate Analysis on **NAME_CONTRACT_STATUS**




**INFERENCES:**

- Maximum loans has been approved by the banks(approx.=78%)
- Canceled loans percentage is very less approx. 1%
- Loans Refused percentage is approx. 20%
- There are approx. 2% loans are unused offer.

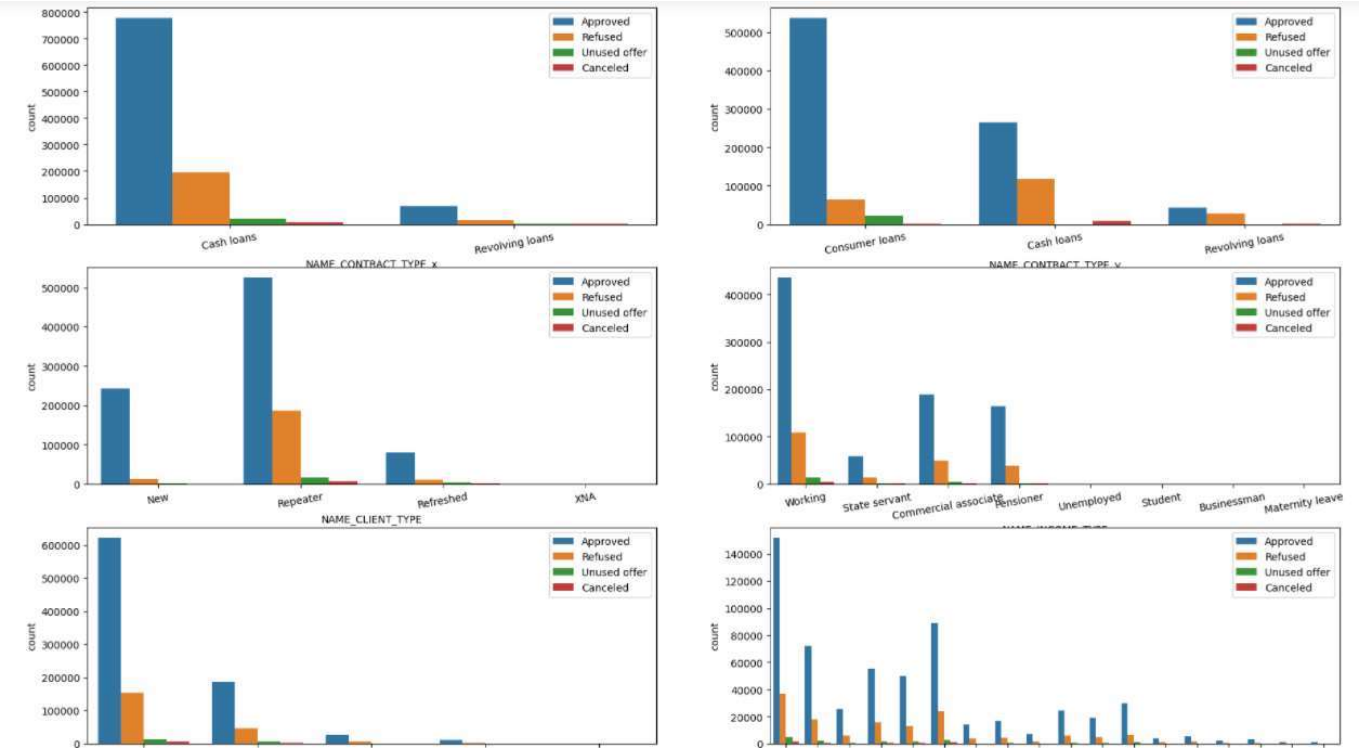
# Bi-Variate Analysis on **NAME_CONTRACT_STATUS VS NAME_CLIENT_TYPE**




**INFERENCES:**

- The loan approval and refused rate for the repeaters is much higher than any other client types
- Loan refused rate for new clients is almost similar like the refreshed clients
- The amount of unused offer and canceled loans are least.
- There are approx. 300k loans are being approved for the new clients which signifies that the banks like to give the loans to the new clients.

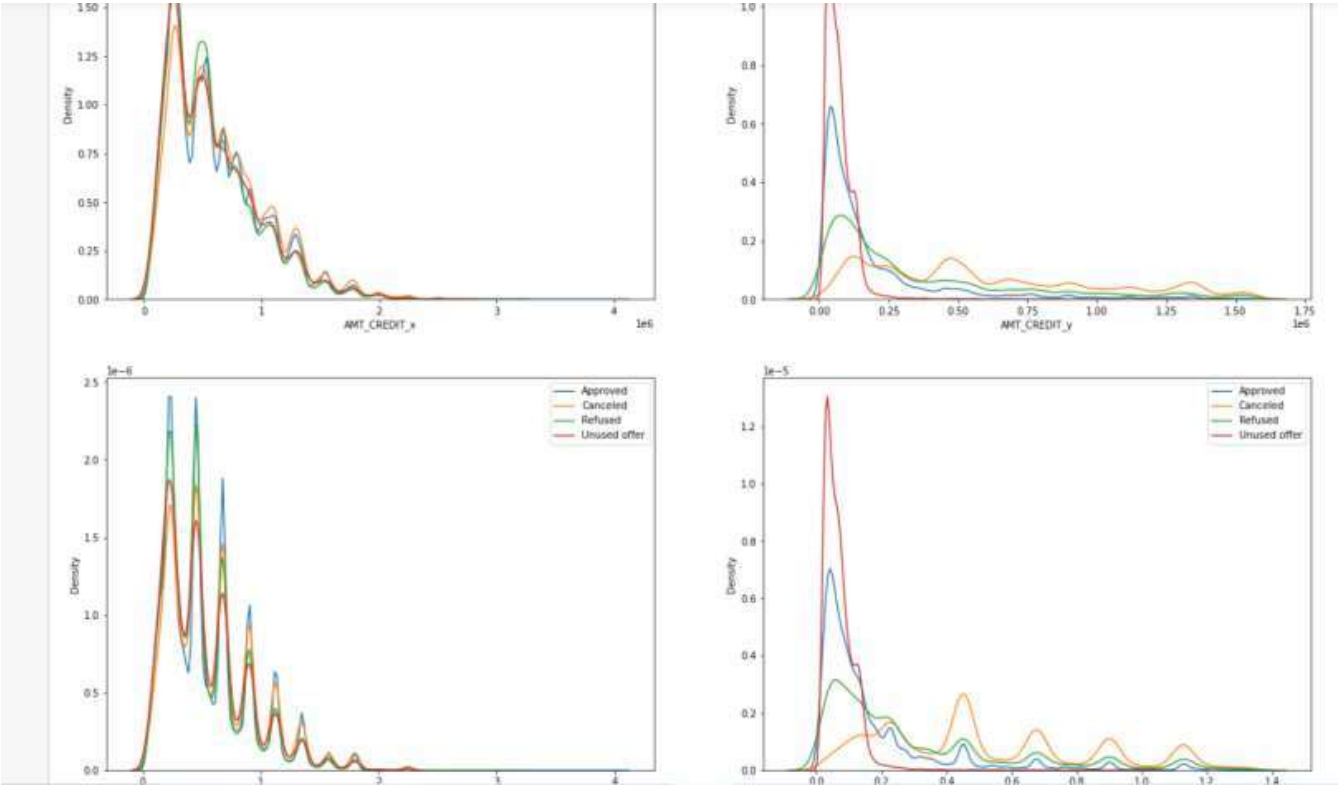
# Univariate Analysis Merged DataFrame (Categorical Columns)




**INFERENCES:**

- Loan approval rates for **Consumer Loans** is much higher than any other loan.
- Banks like to give loans to the **Repeaters**.
- People with **Secondary Education or more** receives loan approval easily.
- Occupation_type **Laborers** get the more loans then others.
- **Working** class people receives more loan approvals than any other Income_type

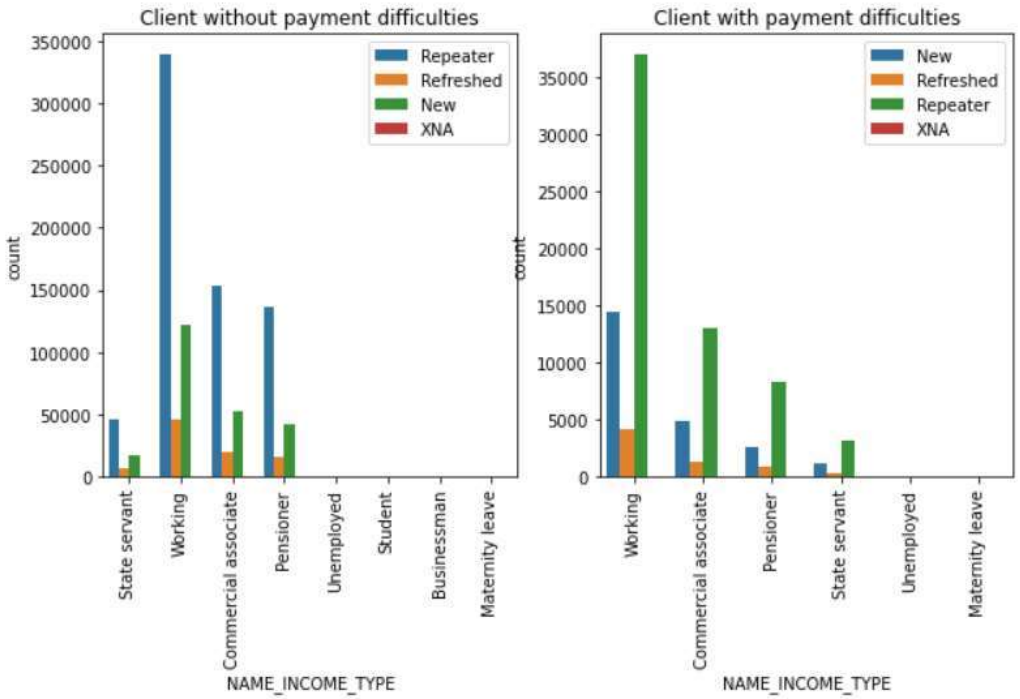
# Univariate Analysis merged DataFrame(Numerical Columns)




**INFERENCES:**

- Loan cancellation rate is higher for a person if he has less number of child or no child.
- Loan approval rate is higher for the family which has family member more than 2.
- Previously bank has more **AMT_CREDIT** for unused offers but now it has more **AMT_CREDIT** for approved loans.
- Loan approval rate is high for the loans which has **AMT_GOODS_PRICE** less than 1 lacs.
- **AMT_APPLICATION** is high for unused offers.

# Bi-Variate Analysis on **NAME_INCOME_TYPE VS NAME_CLIENT_TYPE**




**INFERENCES:**

- For analyzing **NAME_INCOME_TYPE** and **NAME_CLIENT_TYPE** effectively we have divided the dataset into two parts by taking the Target Column as reference variable.

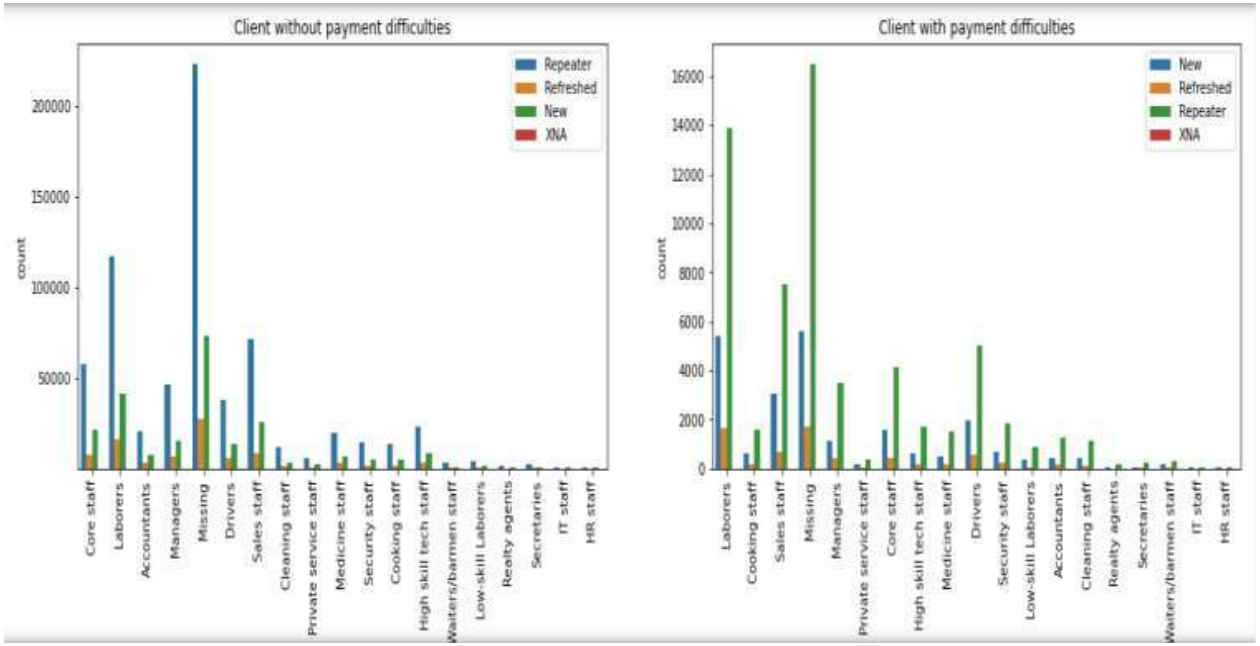
- The divided columns are Client with payment difficulties and Clients without payment difficulties.

- By dividing them into two sets we can conclude these two inferences:-

1. Working Class people take most of the loans.

2. Unemployed, Students, Businessman, Maternity leaves people don't take loans

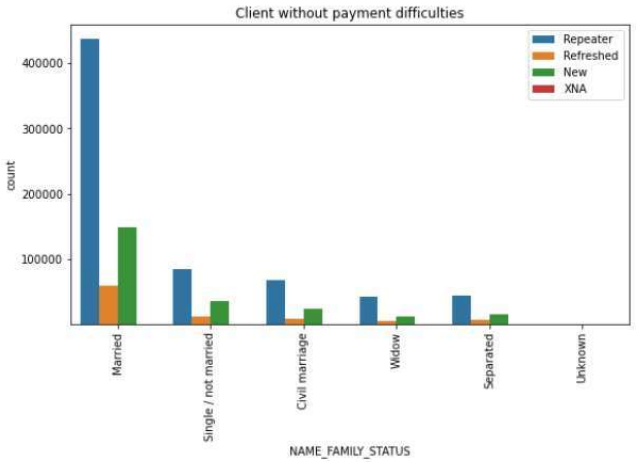
# Bi-Variate Analysis on **OCCUPATION_TYPE VS NAME_CLIENT_TYPE**

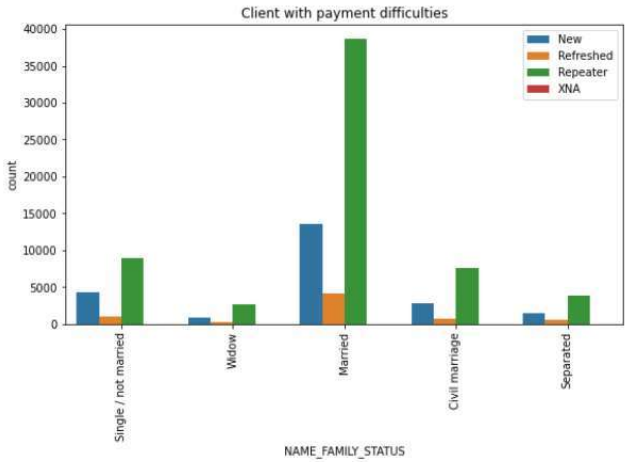



**INFERENCES:**

- Number of Repeaters are very large in both the payment with difficulties and payment without difficulties.
- IT and HR staff people are very less in number which apply for the loans.
- The second top client which face difficulties in paying out the loans is Sales staff.

# Bi-Variate Analysis on **NAME_FAMILY_STATUS VS NAME_CLIENT_TYPE**




**INFERENCES:**

- Married people and the repeaters are the top most category in both the client with payment difficulties and client without payment difficulties.
- Widows and Separated people don't apply much for the loans.
- Single people is the second most category which apply for the loans.

# Conclusion

After performing the analysis we reach out to below conclusion:

1. Banks should consider these variables as loans predictors:-

a) NAME_FAMILY_STATUS
b) AMT_CREDIT
c) OCCUPATION_TYPE
d) NAME_INCOME_TYPE
e) CNT_FAM_MEMBERS
f) CNT_GOODS_PRICE

2. Banks should focus more on cash loans and revolving loans as the percentage of both the loan types are very less in compare to the consumer loans.
3. Banks should encourage widows and separated people to take the loans as there are very less number of people within these two categories which are applying for the loans.
4. Banks should focus on providing low interest rates to the married and working class people since these two categories are the top most category which is facing the highest payment difficulties.
5. Bank should focus less on the people in the low income range .