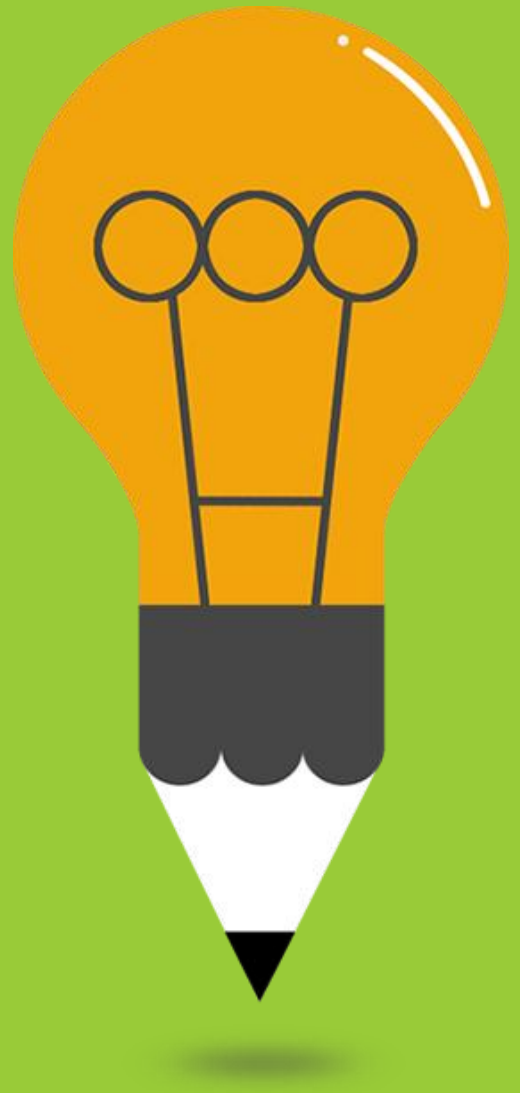
International Institute of Information Technology Bangalore

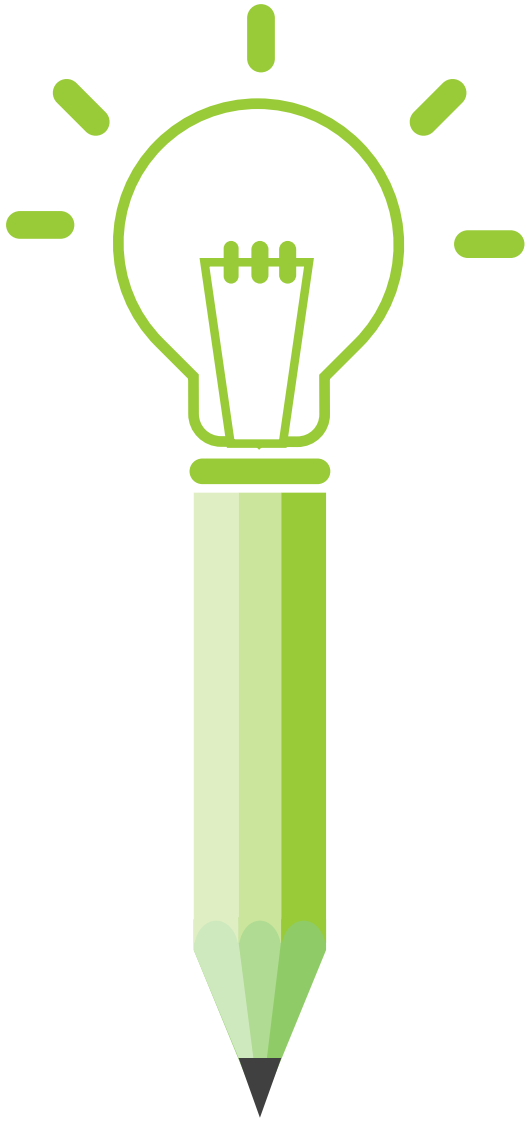
upGrad



EDA Case Study

Presented by -
Prajna Mohanty
Raghav Shankar

Agenda



01 Problem Statement

02 Background

03 Approach

04 Analysis

05 Conclusion

Problem Statements

Business Objective

This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

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:

1. **If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the Company.**
2. **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 below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

1. **TARGET 1 - 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**
2. **TARGET 0 - All other cases: All other cases when the payment is paid on time**

Background

Loans and Type of Loans

A loan is a credit sum issued by a lender to a borrower with an agreement of repaying the credit within a period of time. In most of the cases, the lender will add some interest or processing charges which the borrower should pay along with the principal credit sum.

Types of Loans

There are various kinds of loans such as

- **Consumer Loans :** In order to finance specific type of expenditures like home mortgage, auto loans etc.., creditor issue finance to consumers
- **Cash Loans :** To meet the unexpected expenditures or consumer needs a quick cash, the loans or credit lend is known as Cash Loans.
- **Revolving Loans :** This is a very flexible loan for repaying and re-borrowing the required money within specific limits. Credit cards are the best example of Revolving loans.

Approach

1. **Data Retrieving**
2. **Data Cleaning**
   a) Handling Incorrect Data and Missing Data
   b) Binning
   c) Checking Data Imbalance in target
3. **Data Analysis**

Univariate Analysis	
Gender Distribution	Education Type
Family status of applicants	Housing Type
Type of Loans	Number of family members
Occupation Type	Number of children
Applicant's status on Owning Flat	Credit Amount Distribution
Applicant's status on Owning Car	Annuity Amount Distribution
Applicant's Suite Type	Age Distribution
Applicant's Income Type	Client Income distribution
Applicant’s Total Income	Client types
Channel Type	Contract Status

Bivariate Analysis
Here we have analyzed the distribution of all the above said parameters with the variable `Target` to get a picture on the client status who are having payment difficulties and who are not having payment difficulties.

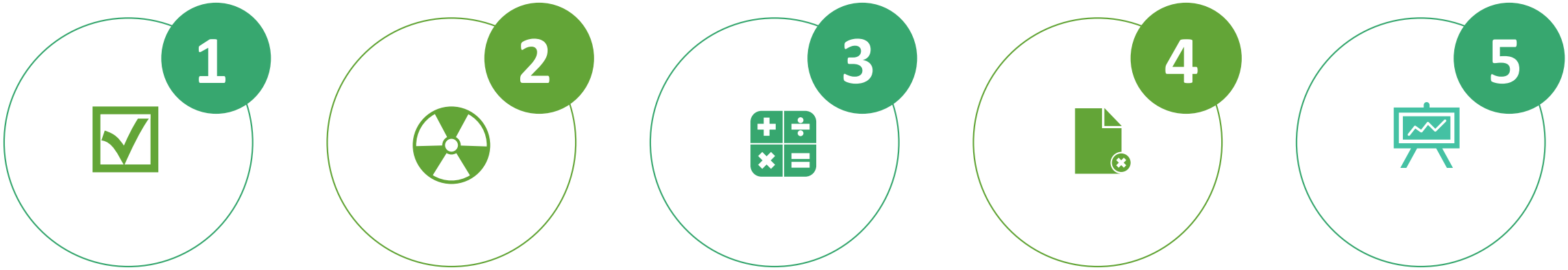
Approach

Analysis

Multivariate and Bivariate Analysis

Exploratory Data Analysis

Major Steps Performed In EDA



Data Importing

Started our EDA by importing the 'application_data.csv' file. We ensured appropriate libraries are imported – Panda, NumPy, Matplotlib.pyplot, Seaborn etc

Data Structure

Basic check on rows, columns and other routine check of the data

Data Quality

Data type check and change the data type, column check for outliers, binning of continuous variable like salary, age etc

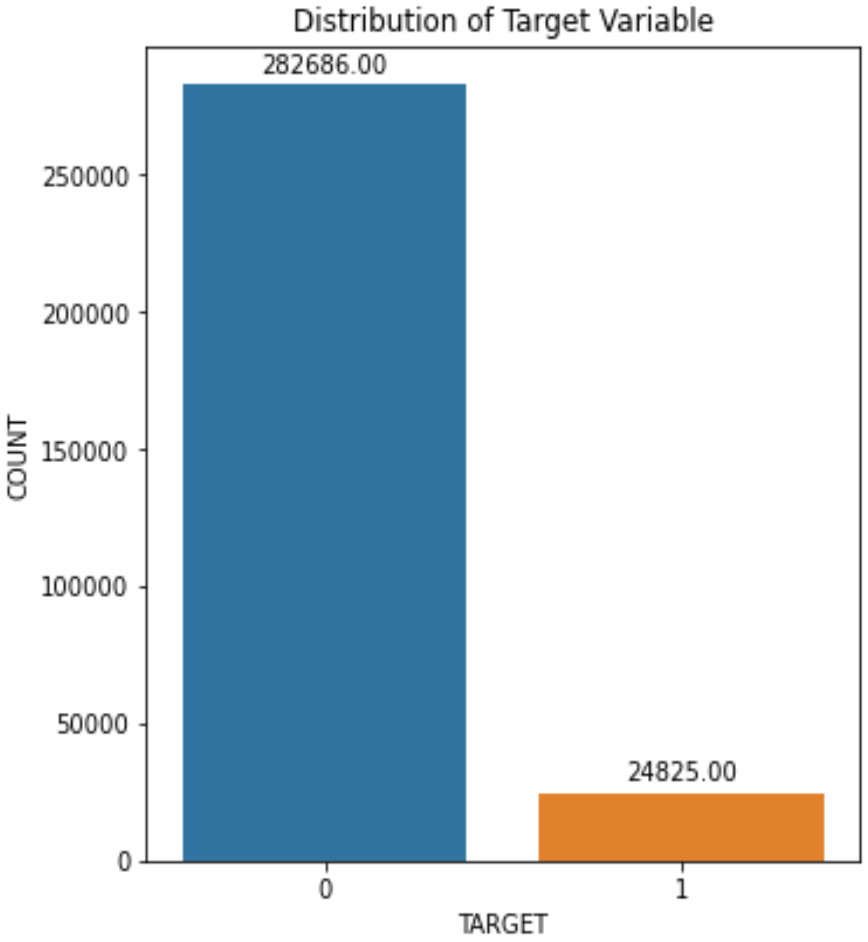
Data Missing Value

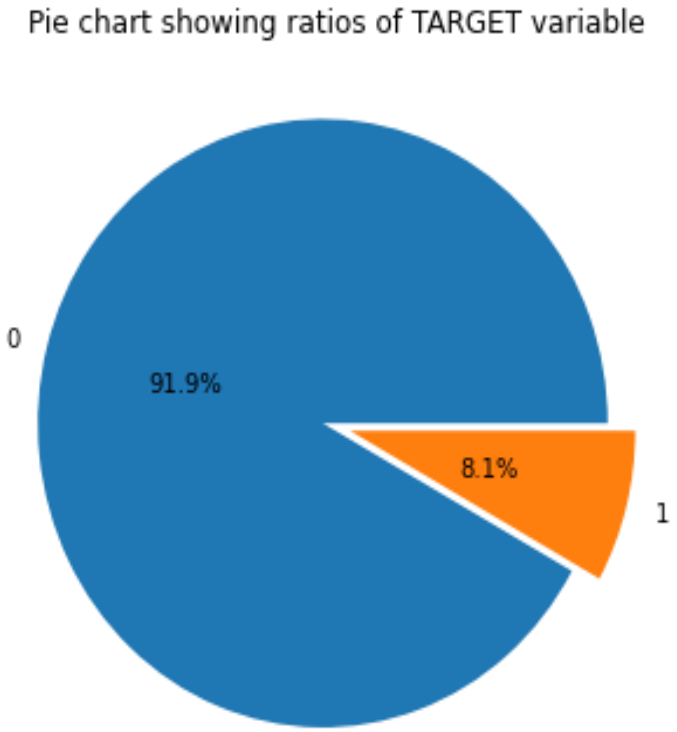
Finding % of missing values, remove columns with high missing values Imputed missing values with median

Analysis

Checked on imbalance % Divided data into tow set, univariate and bivariate analysis and correlation

CHECKING DATA IMBALANCE IN TARGET

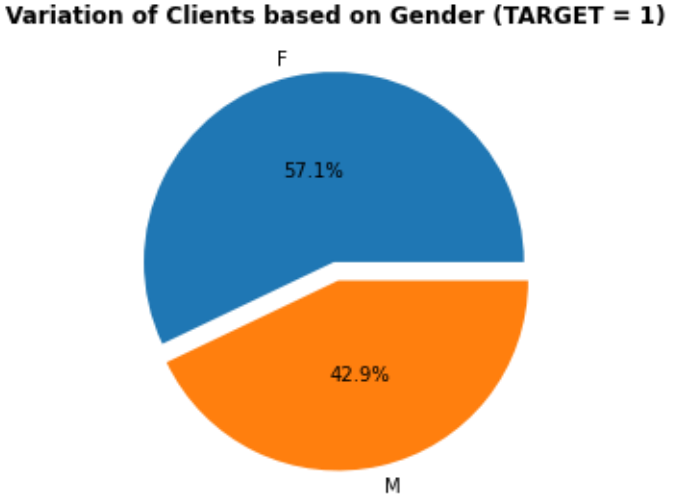


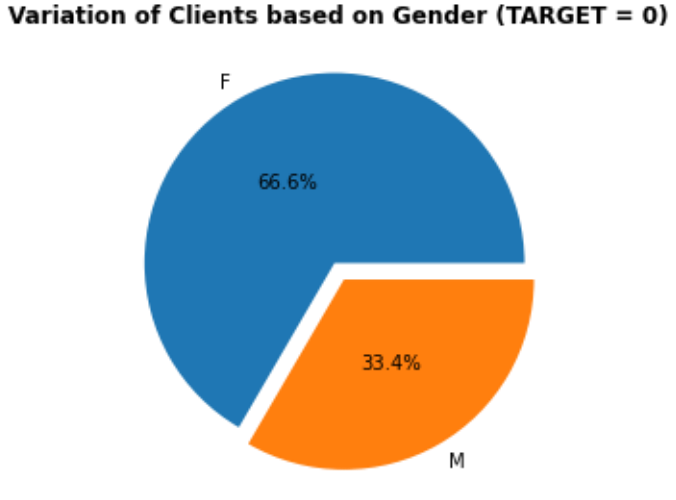


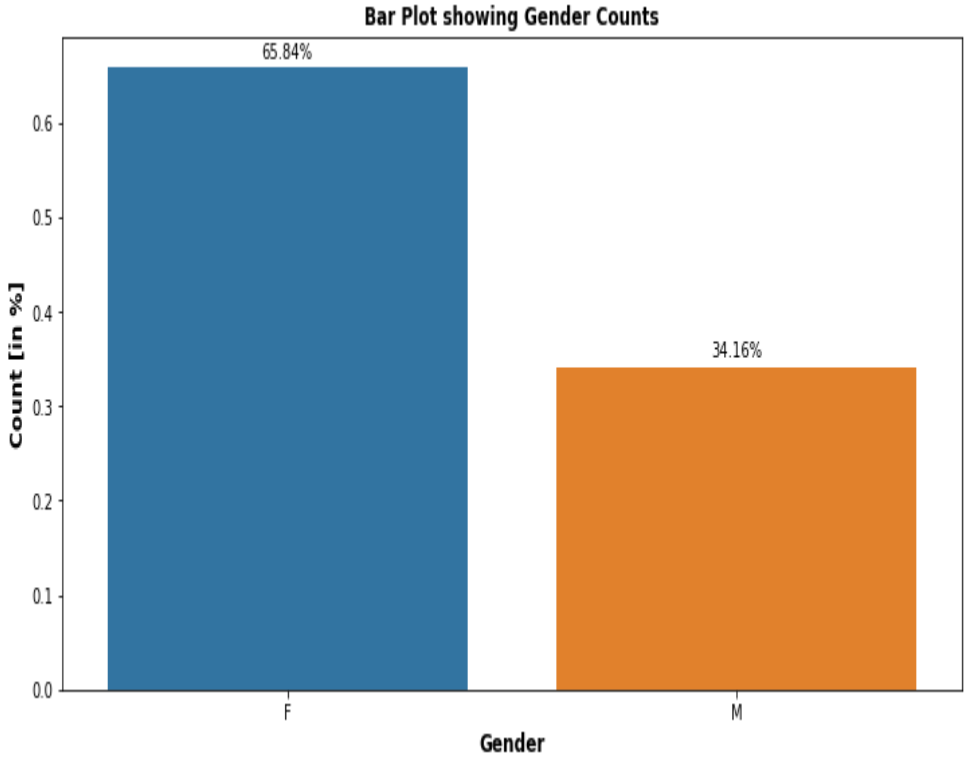
Inference : Here we can see that clients who are repaying the loan without getting default is almost 92% and only 8% of given population is unable to repay loans properly. It explicitly shows that there is data imbalance in target variable.

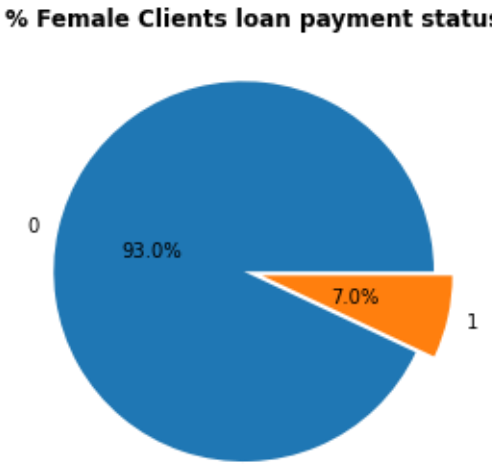
Univariate and Bivariate Analysis

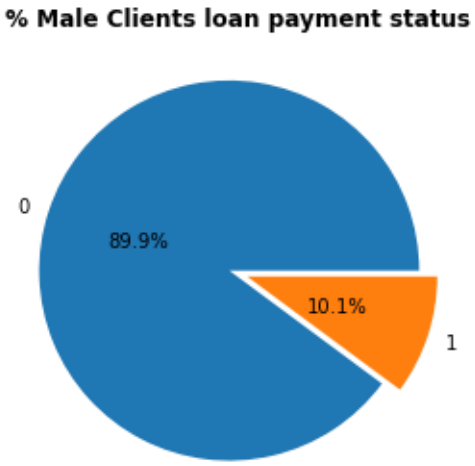
GENDER VS TARGET





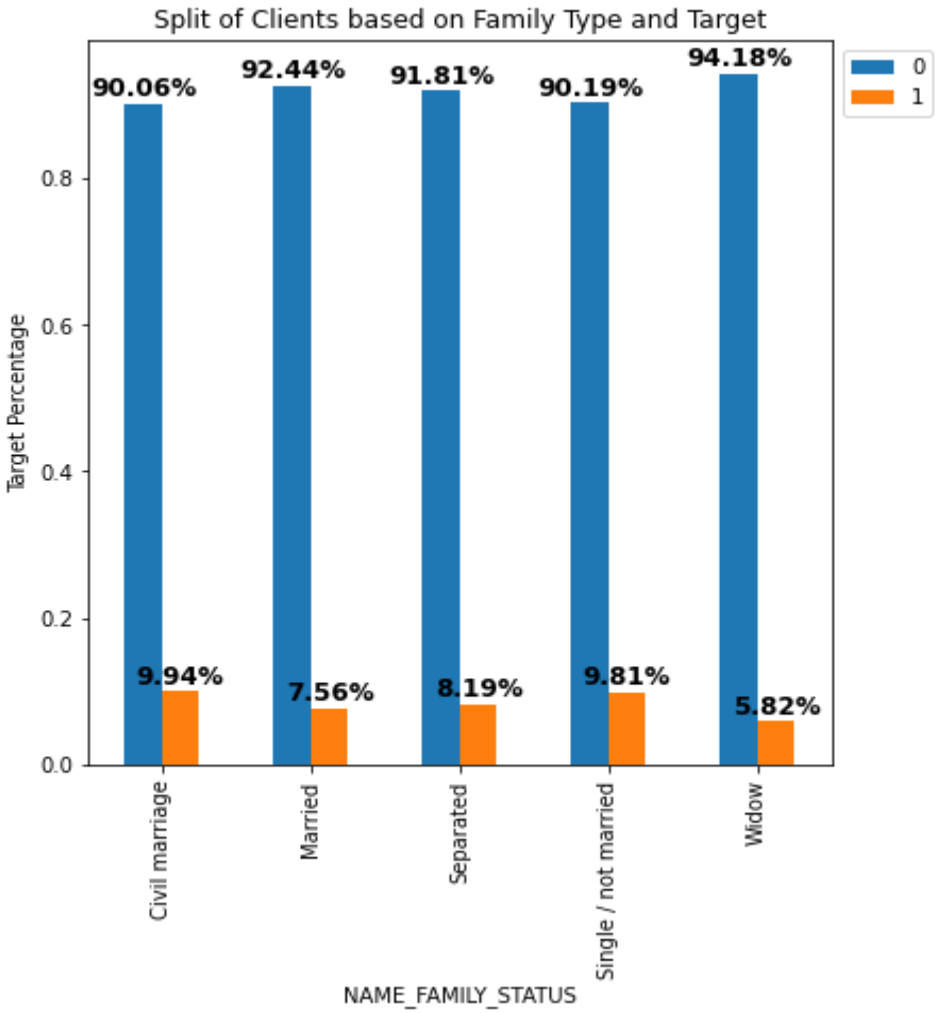






Inference : It is clear from above relation that female clients have applied for loan more than male clients. Also, with repayment status, even Female Clients are twice than males, male clients tend to have higher percentage (10.1% approx.) that they are unable to repay loan.

Family Status VS TARGET

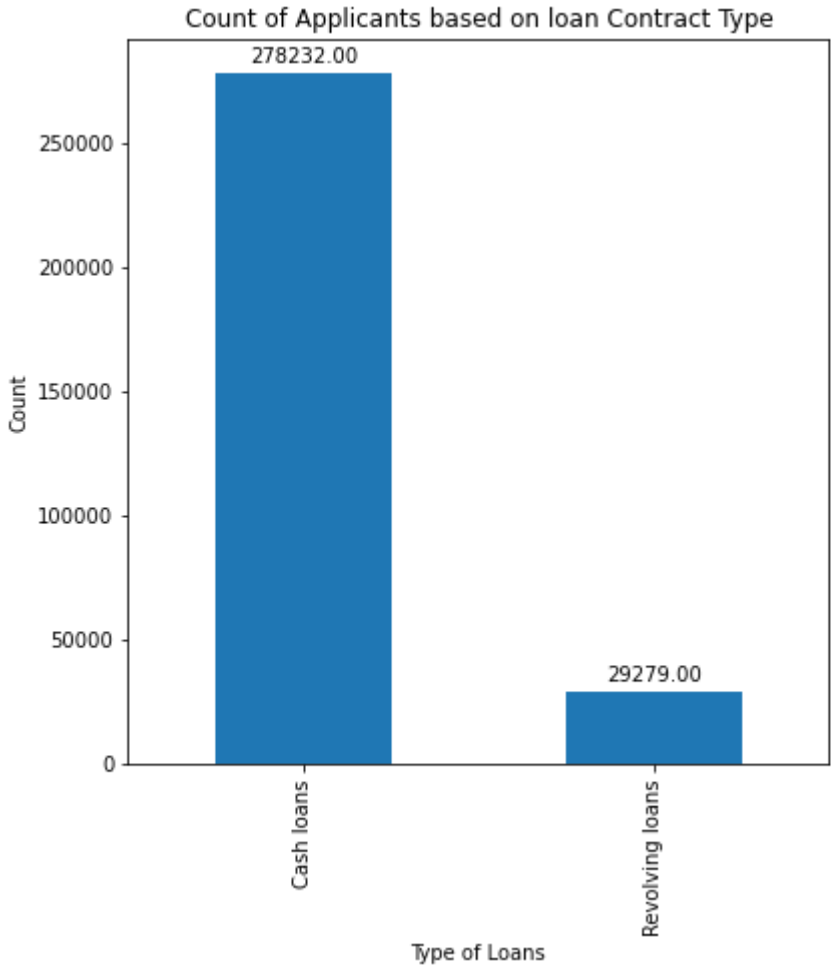


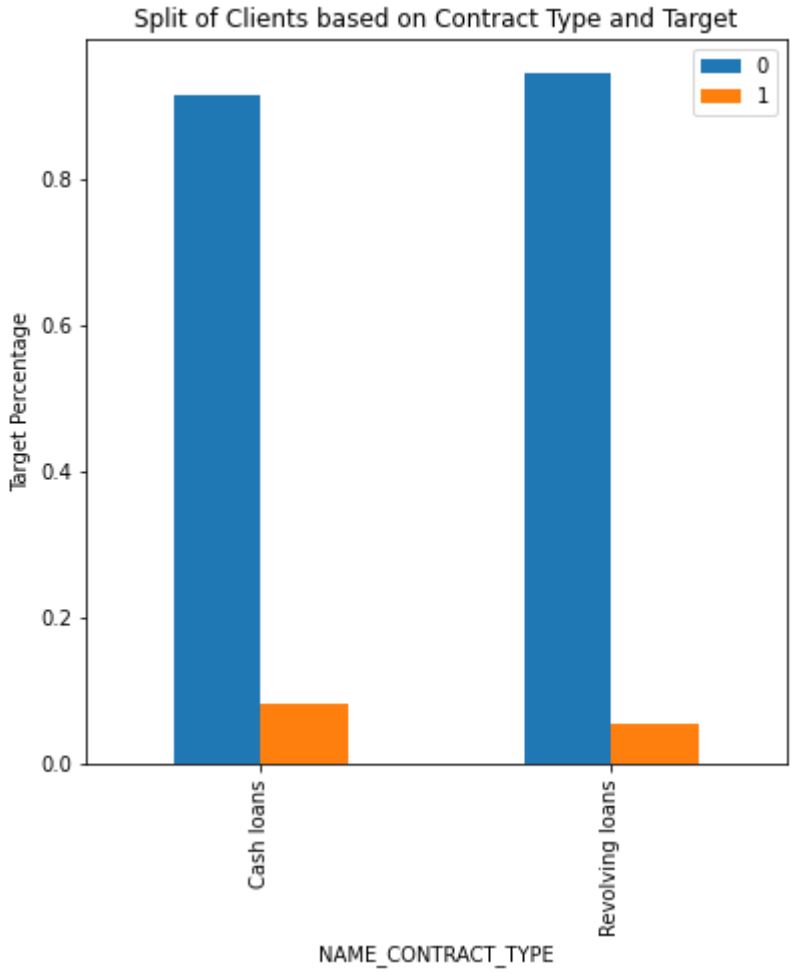
Inference: A very interesting fact to notice here is clients with Married status applied loan the most. The count of Married Clients is far away higher than other Clients.

Also, It is obvious that percentage of Clients unable to pay loan is approximately closer (9.9% approx.) for Clients in Family status Civil Marriage and Single / not married category. Also, these categories have higher percentage of people under non-repayment of loan category.

It is noticeable that Client which is the highest loan borrower i.e, Married people, have little lower percentage of non-repayment of loan than Civil Marriage and Single / non-married people

Contract Types VS TARGET

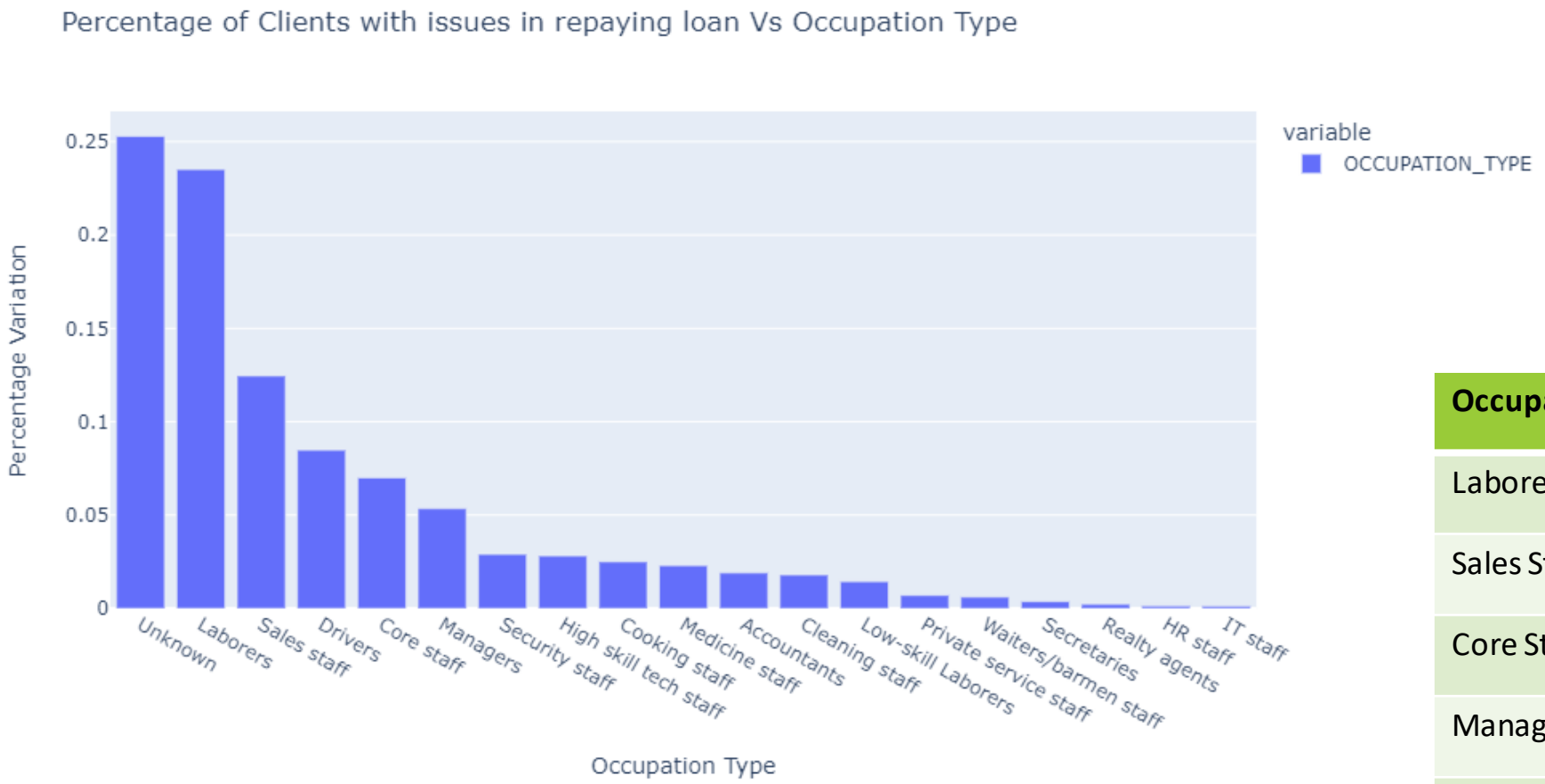




Inference :

- It is clear from the above graphs that Cash loans are predominant in nature than Revolving Loans. And People who are struggling to repay loans also belong to cash loans the higher.
- But the difference in average of people struggling to repay cash loans that of revolving loans are nearer to each other (only 3%) i.e., 8.3% - 5.4% = 2.9%
- So, we can find that people struggles the most to repay Revolving loans.

Occupation VS TARGET



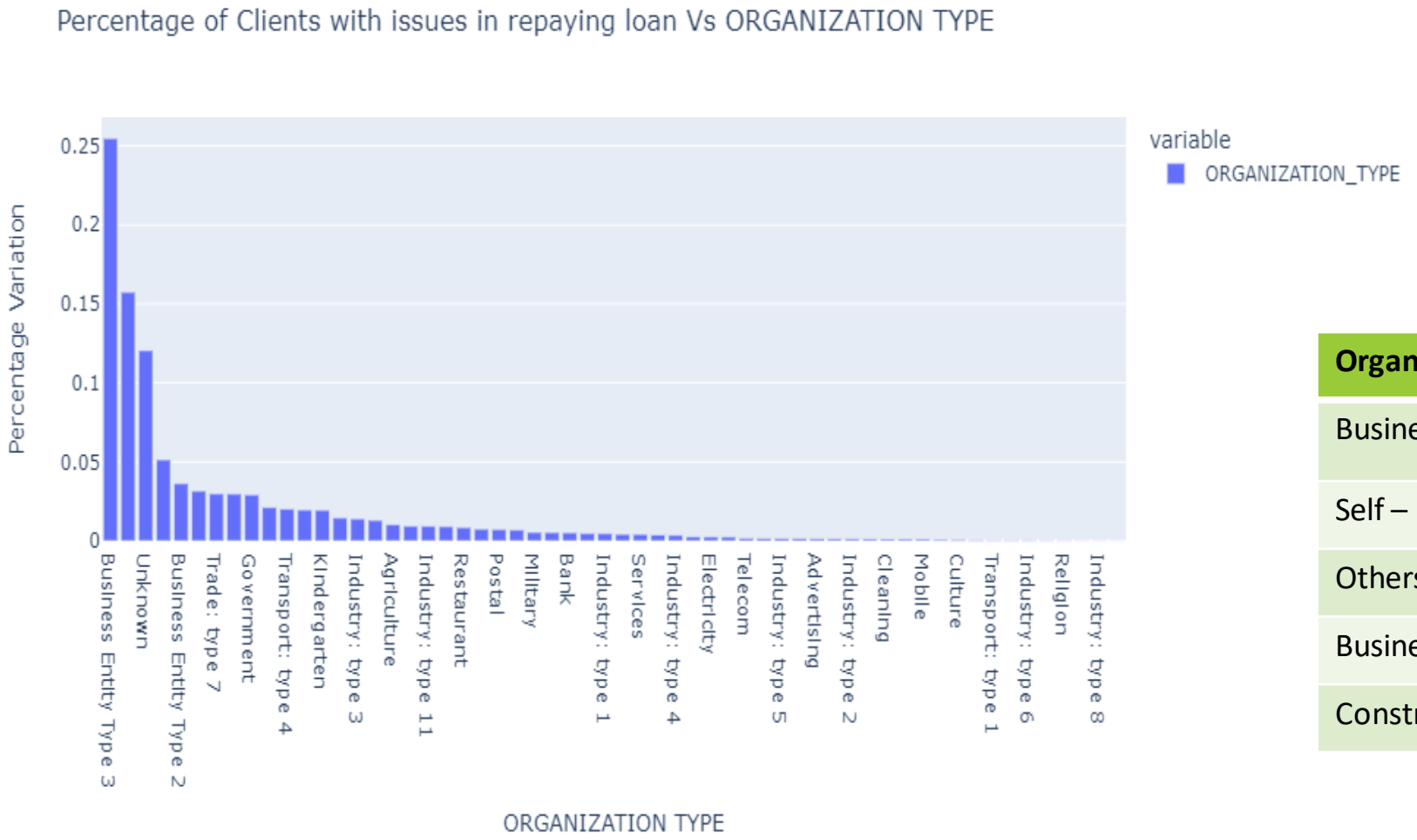
Inference:

Data on Type of Employees and Count and Percentage of TARGET = 1

Neglecting the category Unknown
Top 5 Occupation Types

Occupation	Count	Percent
Laborers	55K	23.52%
Sales Staff	32K	12.45%
Core Staff	27K	7.00%
Managers	21K	5.34%
Drivers	18.6K	8.48%

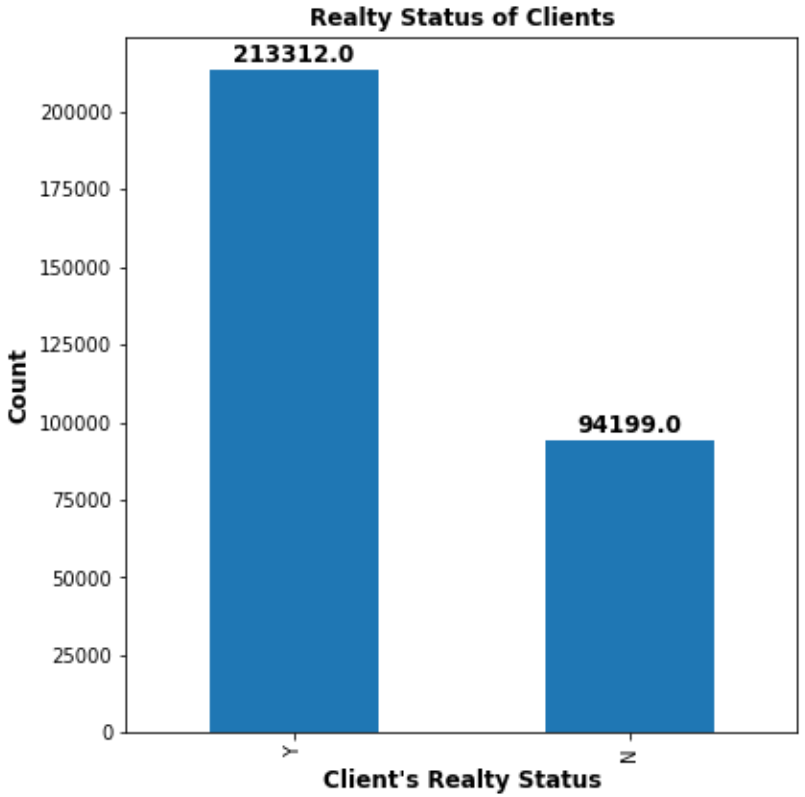
Organization VS TARGET

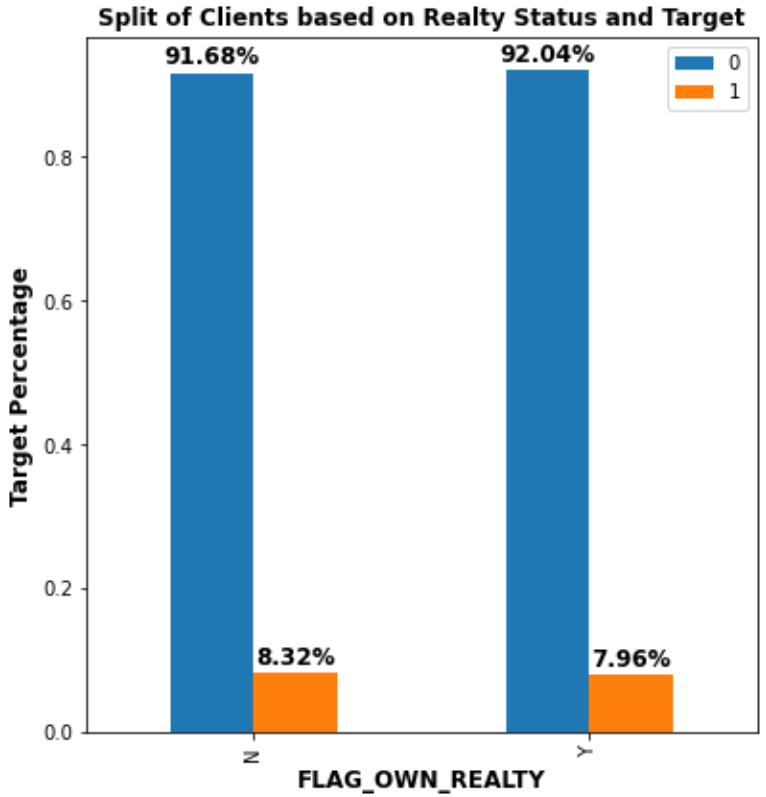


Inference : Below is the pattern of defaulters of loan in descending order based on Organization Type
Top 5 Defaulters (Except Unknown):

Organization Type	Percent
Business Entity Type 3	25.47%
Self – Employed	15.74%
Others	5.14%
Business Entity Type 2	3.62%
Construction	3.162%

TARGET VS CLIENT OWNING FLAT

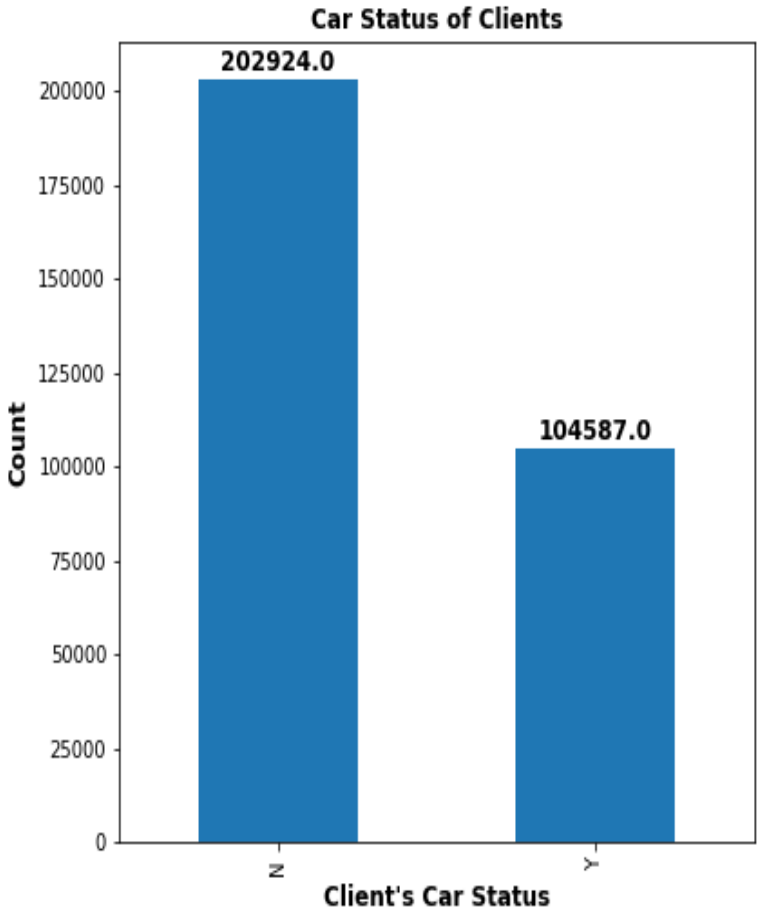


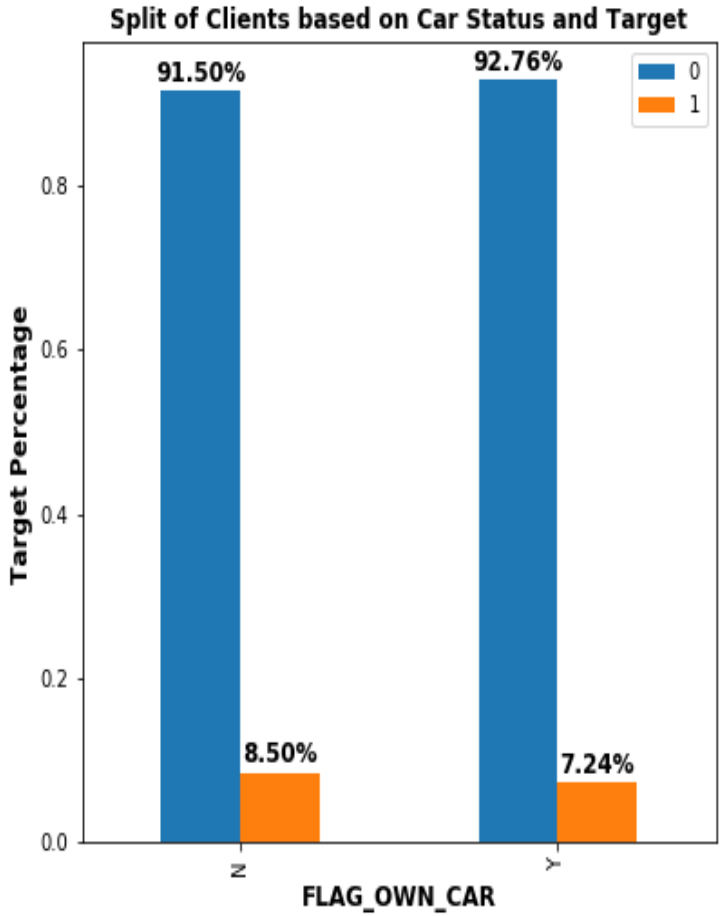


Inference :

- The above plot depicts that the count of Clients owning a flat or house who applies for loan is far higher than those who are not owning any house or flat.
- In Contrast, we can see from the plot 2 that, irrespective of client owning house or flat, the average of client's facing issues in repaying loan (Target = 1) is approximately nearer to each other (8% approx).

Applicants Owning Car Vs Target

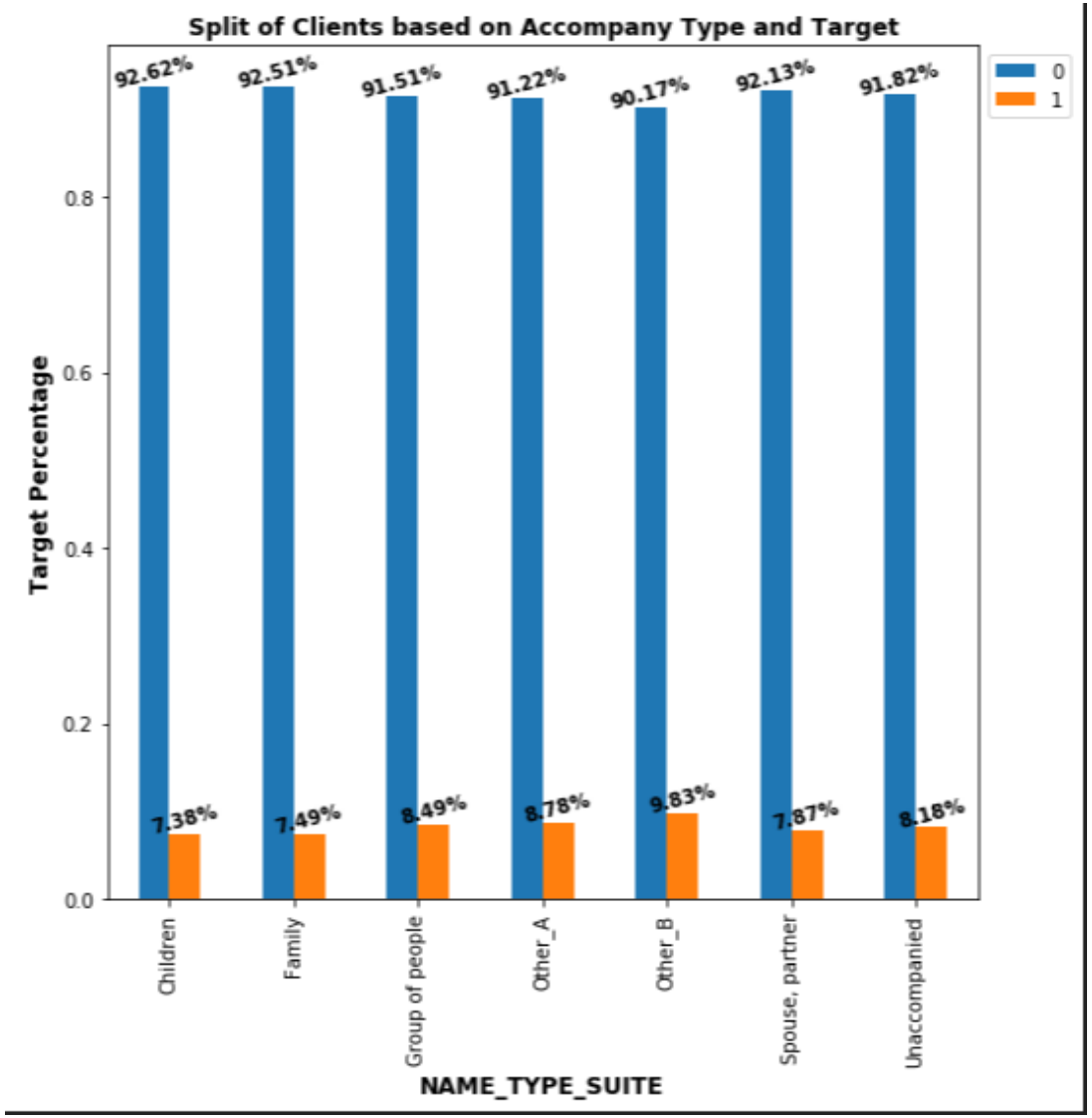




Inference:

- Comparing to Clients owning Flat, number of clients owning car is somewhat lesser.
- Also, there is a slight variation in percentage of Clients not owning car who are unable to pay loans (8.5% approx) than those clients owning car who are unable to repay loans (7.2%)

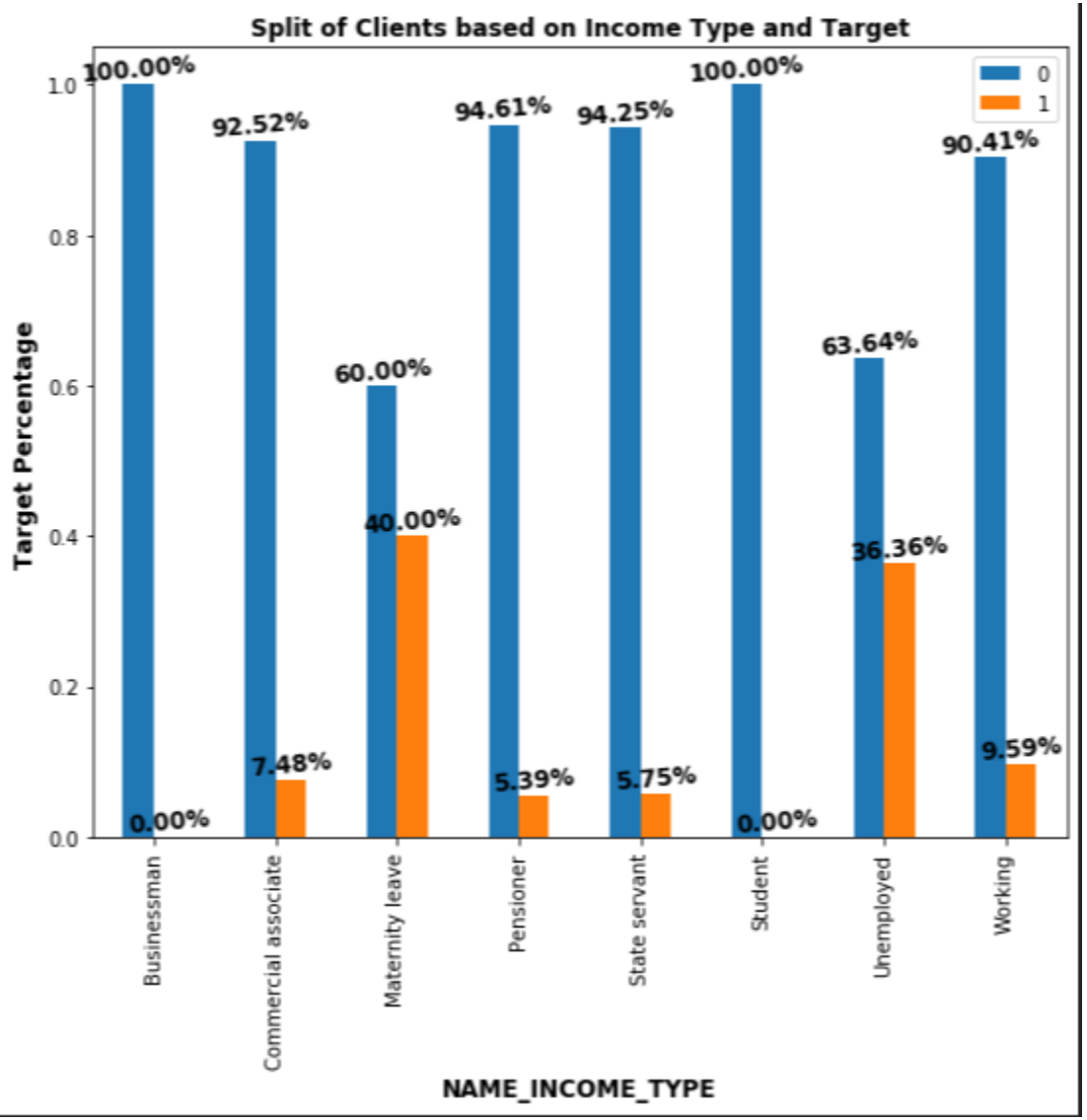
Applicant's Suite Type Vs Target



Inference :

- The fact seen from the above plot is that there is no huge difference in percentage of people who are not repaying loans with Accompany status as unaccompanied with other Accompany types.
- This shows that Accompany status have no big impact on people not repaying loans

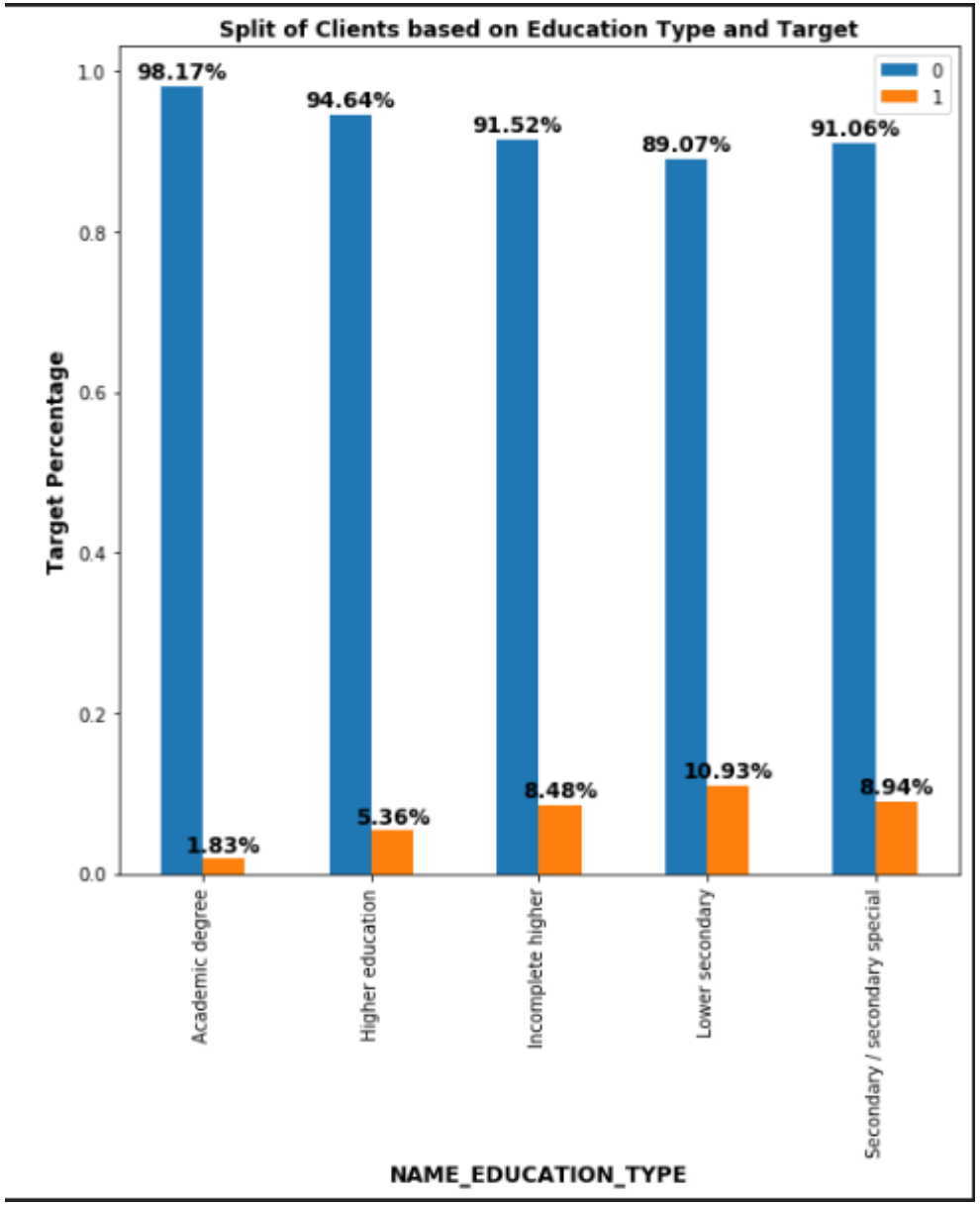
Applicant's Income Type Vs Target



Inference:

- It is quite interesting to see loan approved for Unemployed Clients too.
- The successful repayment status of loans based on Income type is following below order,
- Businessman, Student > Pensioner > State Servant > Commercial associate > Working > Unemployed > Maternity Leave
- It is obvious here that banks/similar companies are strict to approve loans to Clients in Income types Unemployed, Student, Businessman, Maternity leave
- Also, 100% Clients in Income types Businessman and Student are paying loans in time.
- Nearly 40% of Clients coming under Maternity Leave and Unemployed are not repaying loans properly falling under TARGET = 1
- So, we can say from this analysis, Banks can focus more on clients other than those falling under income type Maternity Leave and Unemployed

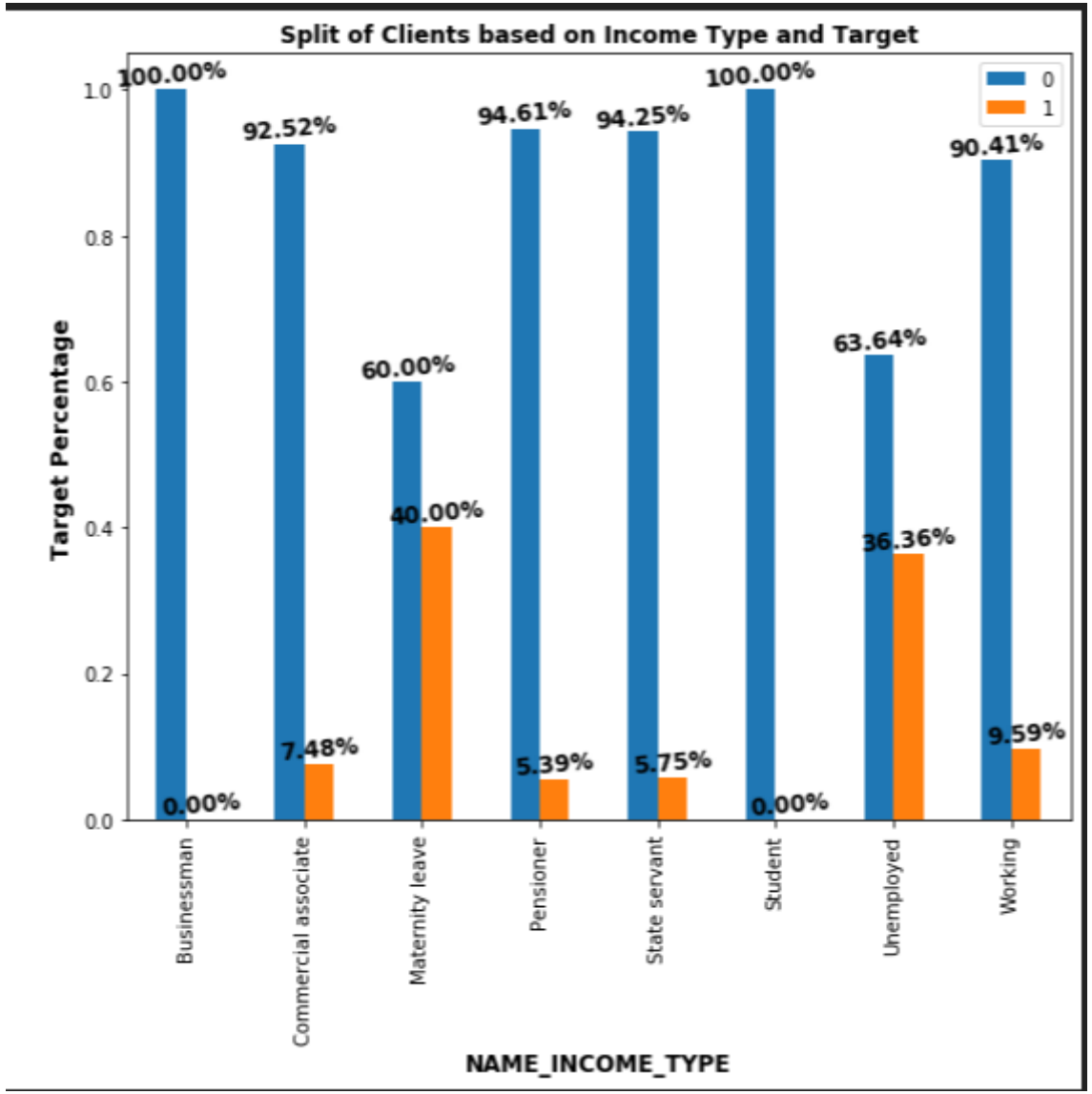
Education Type Vs Target



Inference:

- It is seen from the plot that clients with Secondary or Secondary special education applies for the loan the most and they stands at second position among the categories who are unable to repay loans
- The top most category who fails to repay loan, in other words falling under category TARGET = 1, is Lower Secondary education type people. But these people are the second last category among the count of people categorized based on education type.
- The category pattern with respect to education Type who are unable to repay loans follows the below format:
- Lower Secondary > Secondary / Secondary Special > Incomplete Higher > Higher education > Academic Degree

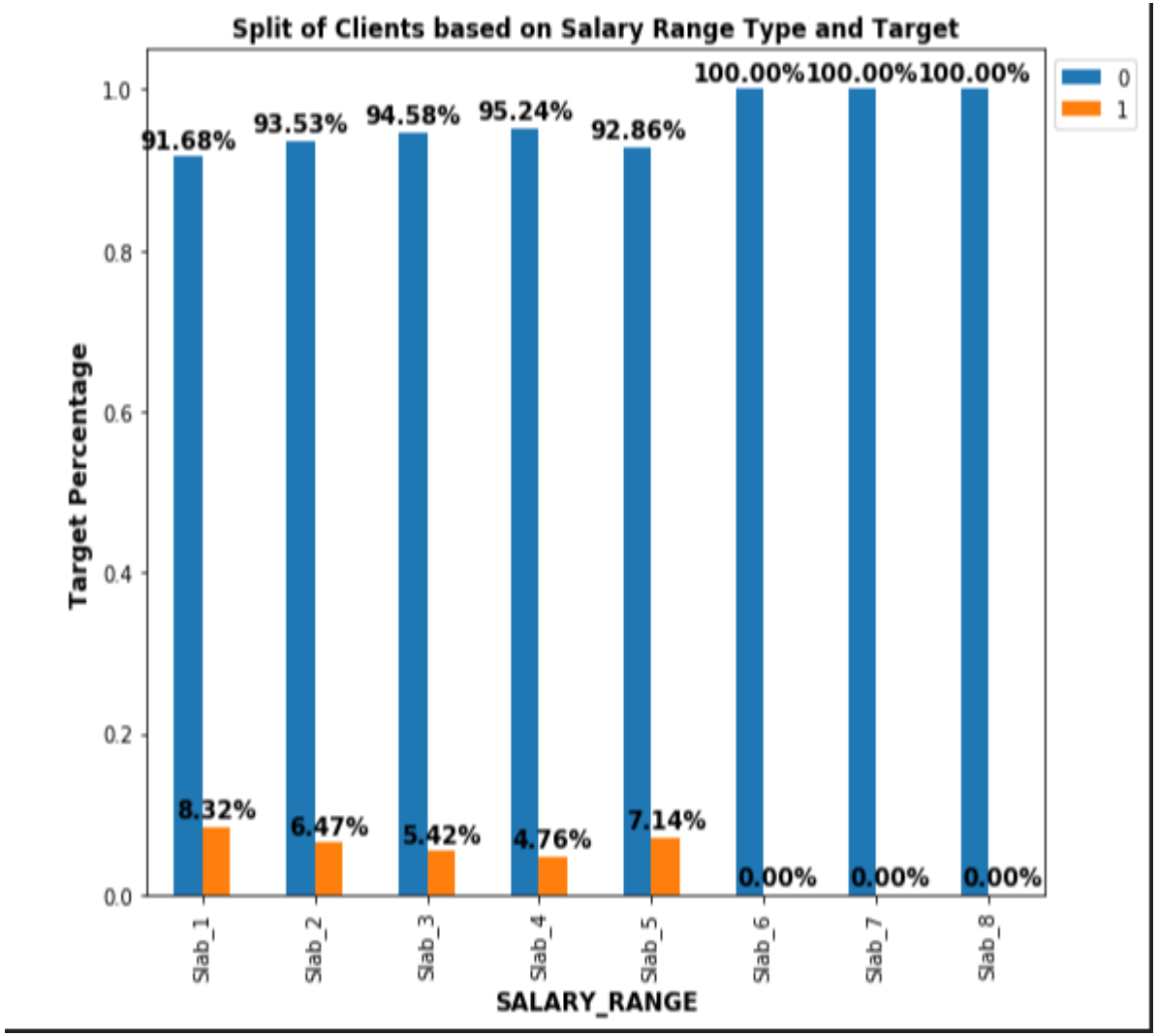
Housing Type Vs Target



Inference

- We can see from above plot that most of the clients who applied loan are staying in House / Apartment.
- Here, we can notice that clients staying in Rented Apartment fails a lot to repay loan (nearly 12.31%). So, we can say that clients are facing challenges in repaying loans and rental to house.
- And we can also infer here Clients staying with Parents are also failing a lot to repay loans (11.7%). We can also say that Clients are having an additional responsibility to take care of their parents which make them unable to repay loans.
- And finally, we can say if Clients having additional expenditures like House rent, Taking care of parents are also affecting the payment of loans or debts

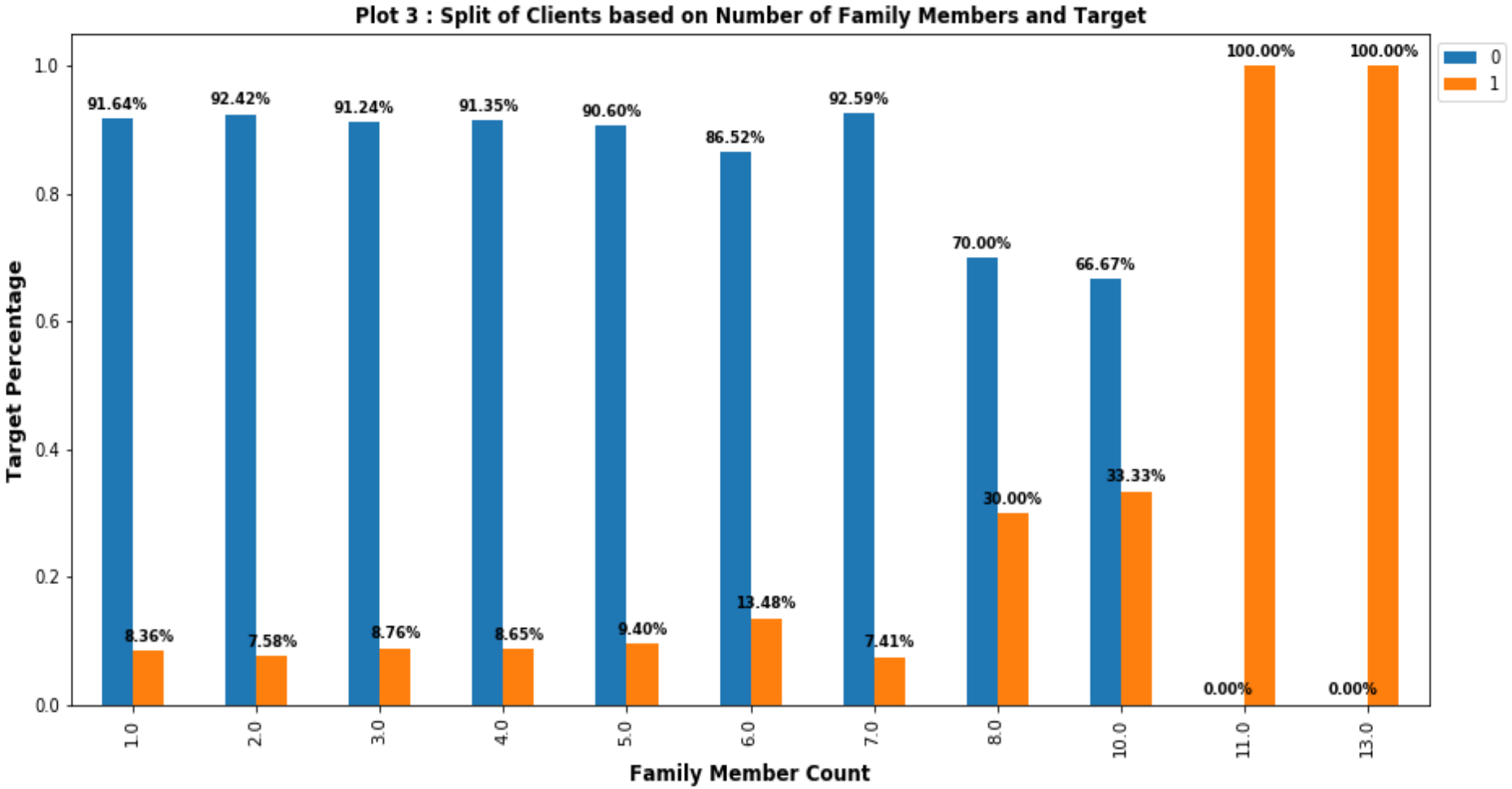
SALARY RANGE Vs TARGET



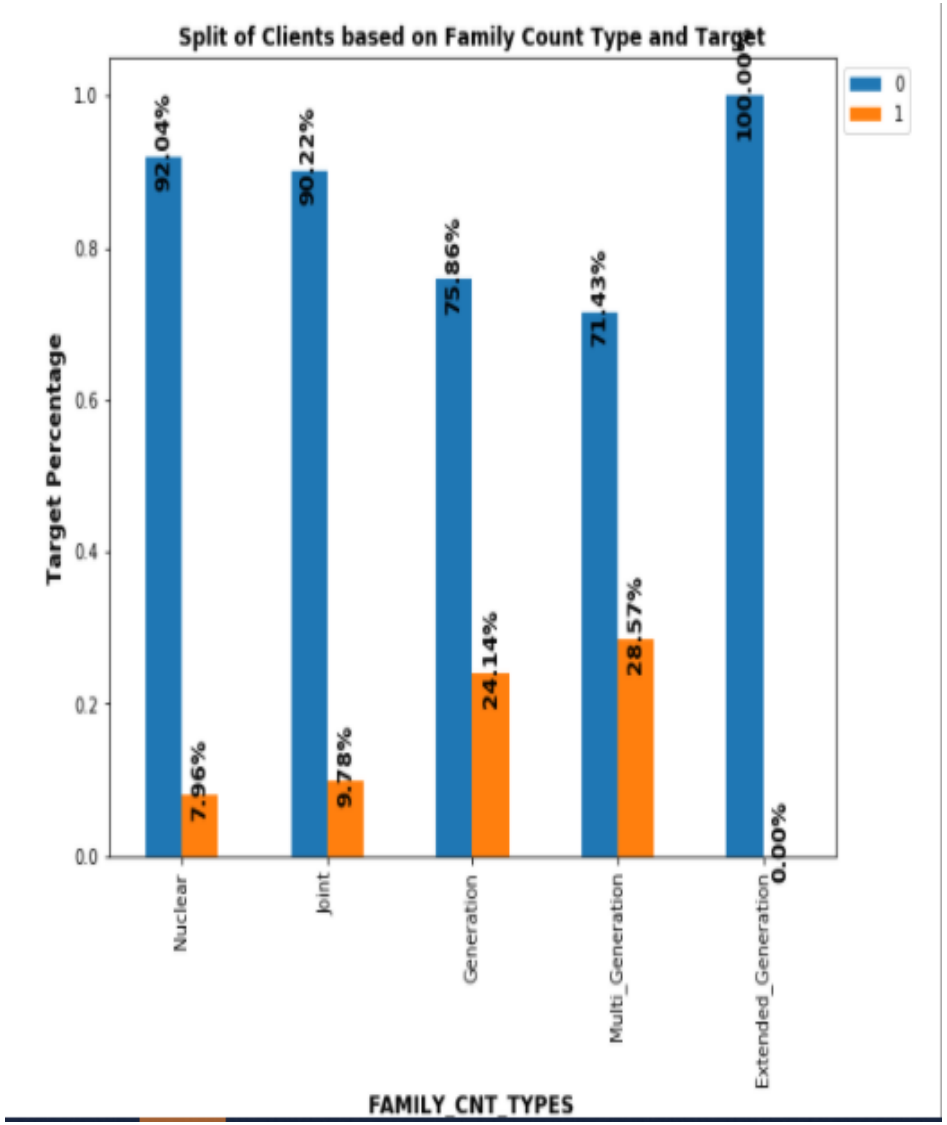
Inference:

- It is clear from above plot that about 99% of clients falls under salary slab 1 and 2 i.e., from clients having income of 0K to 500K.
- Also, the defaulter clients are maximum for Slab 1. And the second maximum defaulters are in slab 5 (Clients with Annual Income 25 LPA to 50 LPA)

Number of Family members Vs Target



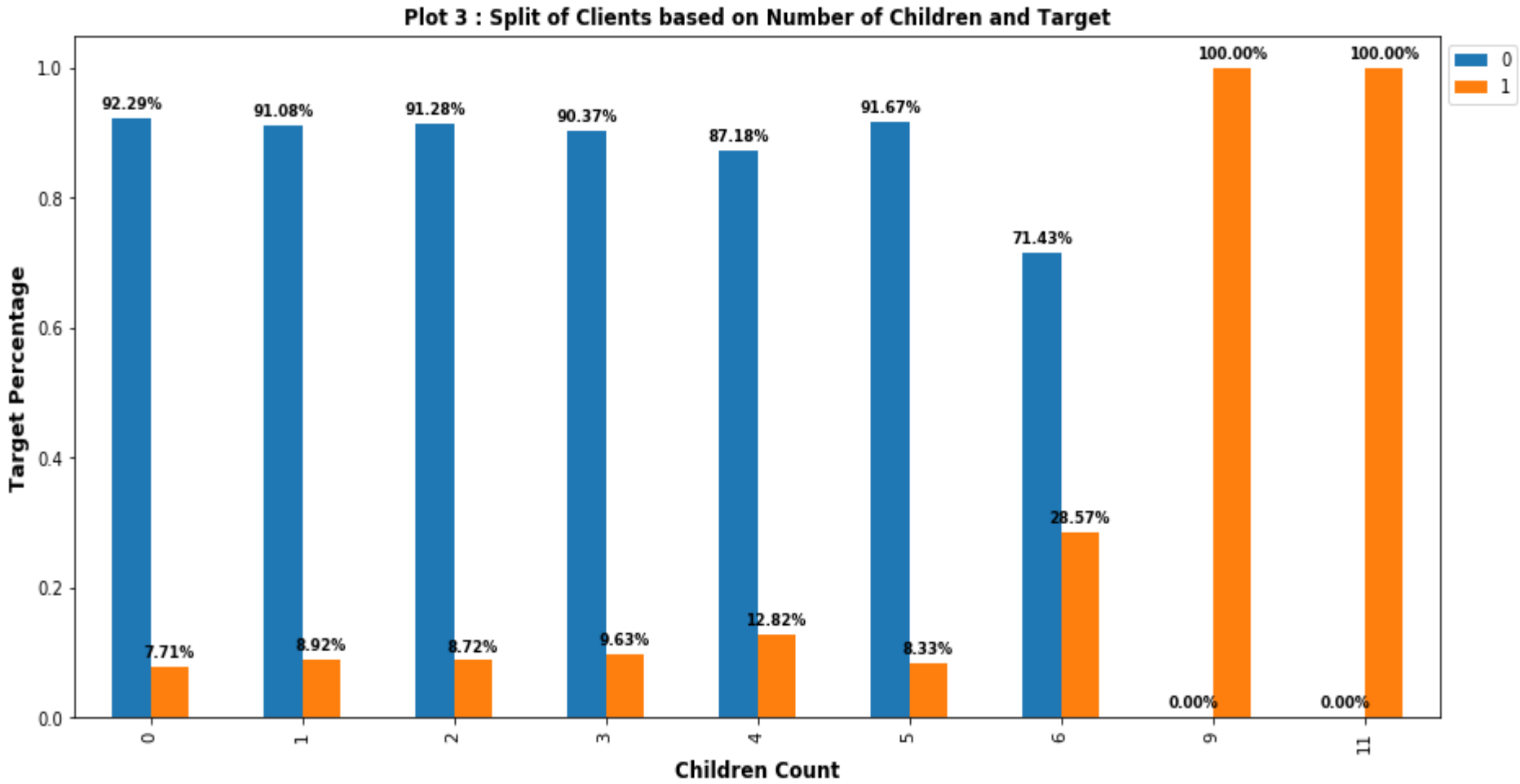
Number of Family members Vs Target



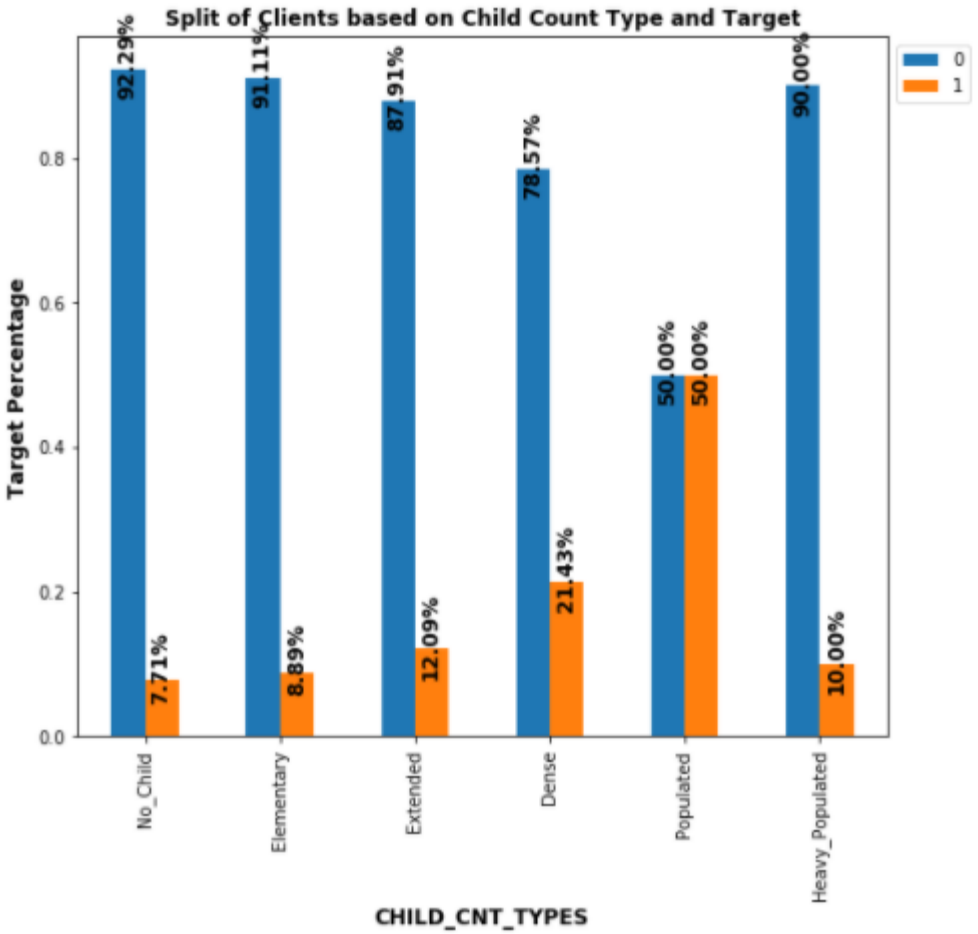
Inference:

- It can be seen that most of the families applying for loan are having family member count of 2.
- Here the visible pattern of higher client counts are falling under family with member count that follows below order
- Family Member Count : 2 > 1 > 3 > 4 > 5
- It is seen from the Plot 3 that Clients with Family member count of 11 and 13 are failing 100% to repay loan but we can infer from Plot 2 that the count of such clients are only 1 each.
- The next category of highest failures is with family member count of 10 and 8. Here too the count of Clients applied for loan is much smaller with 3 and 20 Clients respectively. The former category has 1 client in TARGET = 1 and the latter has 6.
- The majority Clients we need to Focus here is the clients with Family member Count 1 to 5. Neglecting other Categories, we can see that the contrast pattern is followed as above for Clients with TARGET = 1
- i.e., Number of Clients with TARGET = 1 Follows: 2 < 1 < 4 < 3 < 5
- We can say that the percentage of people unable to repay loan is somewhat lesser for Clients with Family Count 1 and 2
- This infer our previous point that the Clients having additional expenditure like With Parents are failing to repay loan. Similarly, Clients managing the expenditures of more number of people in family are unable to repay loan in time

Number of Children Vs Target



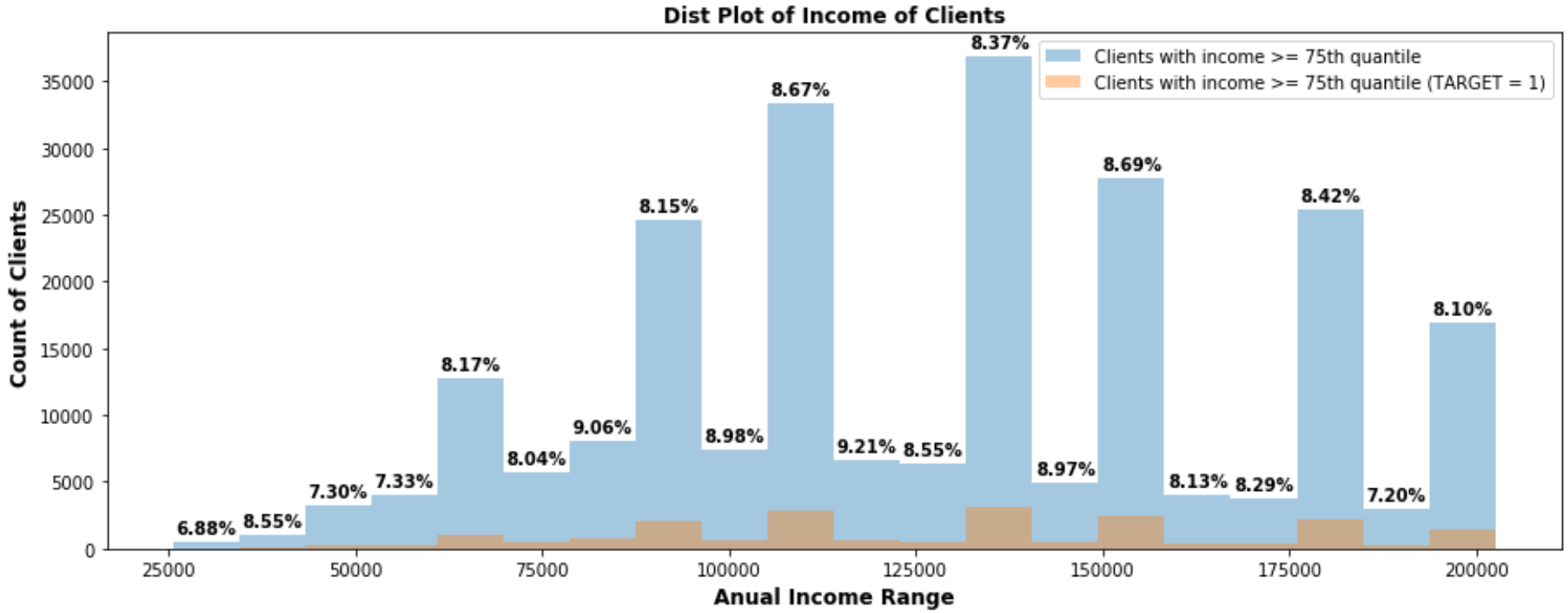
Number of Children Vs Target



Inference:

- We can infer from the plot 1 that maximum number clients applied for loan have no children (215371 nos.). From Plot 3, it is obvious that the clients having Children count of 6 (21 clients in total) fails the most to pay the loan.
- While analyzing the remaining population (from number of children from 0 to 4), where the maximum of Clients falls under, it is seen that Clients inability to repay loan decreases with increase in number of children except for Child count 2.
- This resembles the same scenario as above that if a Client is committed to take care of more number of family members, which in turn increases the client's expenditure, are unable to repay loans properly

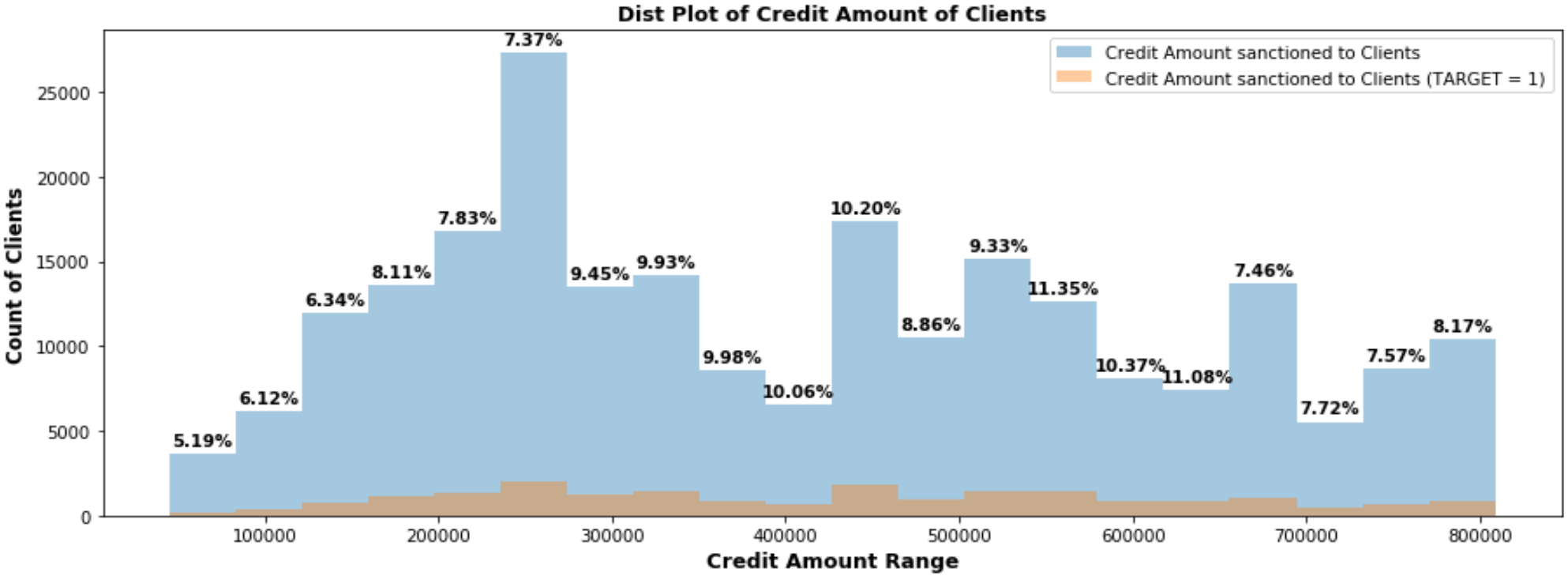
Distribution of Annual Income of Clients Vs TARGET



Inference:

- We can infer here that the most of the clients are falling under income range 130K to 140K.
- But this category is not the first category of people under TARGET = 1.
- Clients with income 120K to 124K are the top most category people facing issues in repaying loans (9.21% approx.) Second most category is of clients with income range 80K to 85K (9.06% approx)
- We see that there are certain distribution of people in all income ranges with minimum about 7% clients are not repaying loans properly

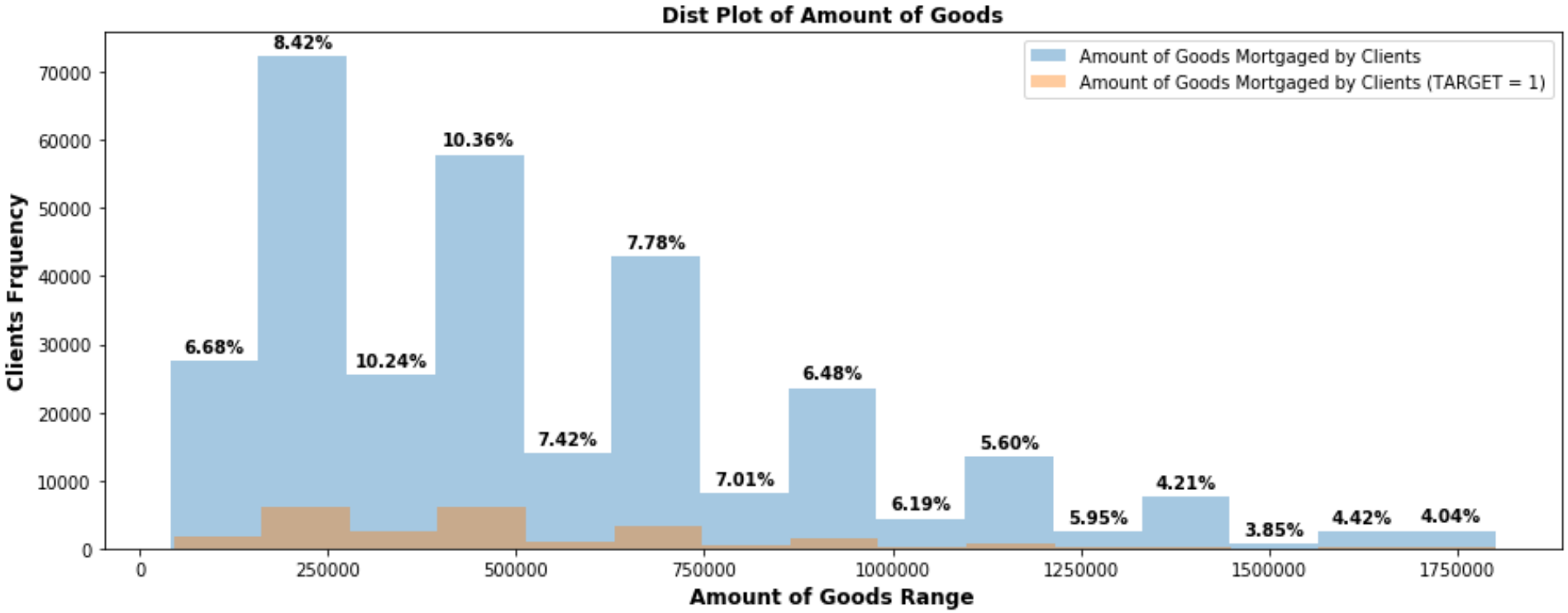
Analyzing Credit Amount with TARGET



Inference:

- In this plot we can see that maximum number of clients falling under TARGET = 1, is in the Credit amount range starting from 2.75 LPA to 6.5 LPA (nearly 9.5% to 11.5%)
- It is seen that large number of clients fall under credit range of 2.5 to 2.75 LPA but these people are not the top most non payers of loan.
- Here we can see that as credit amount increases, from certain value (from 2.75 LPA) to a certain value (6.75 LPA) it seems the clients are facing issues in repaying loans

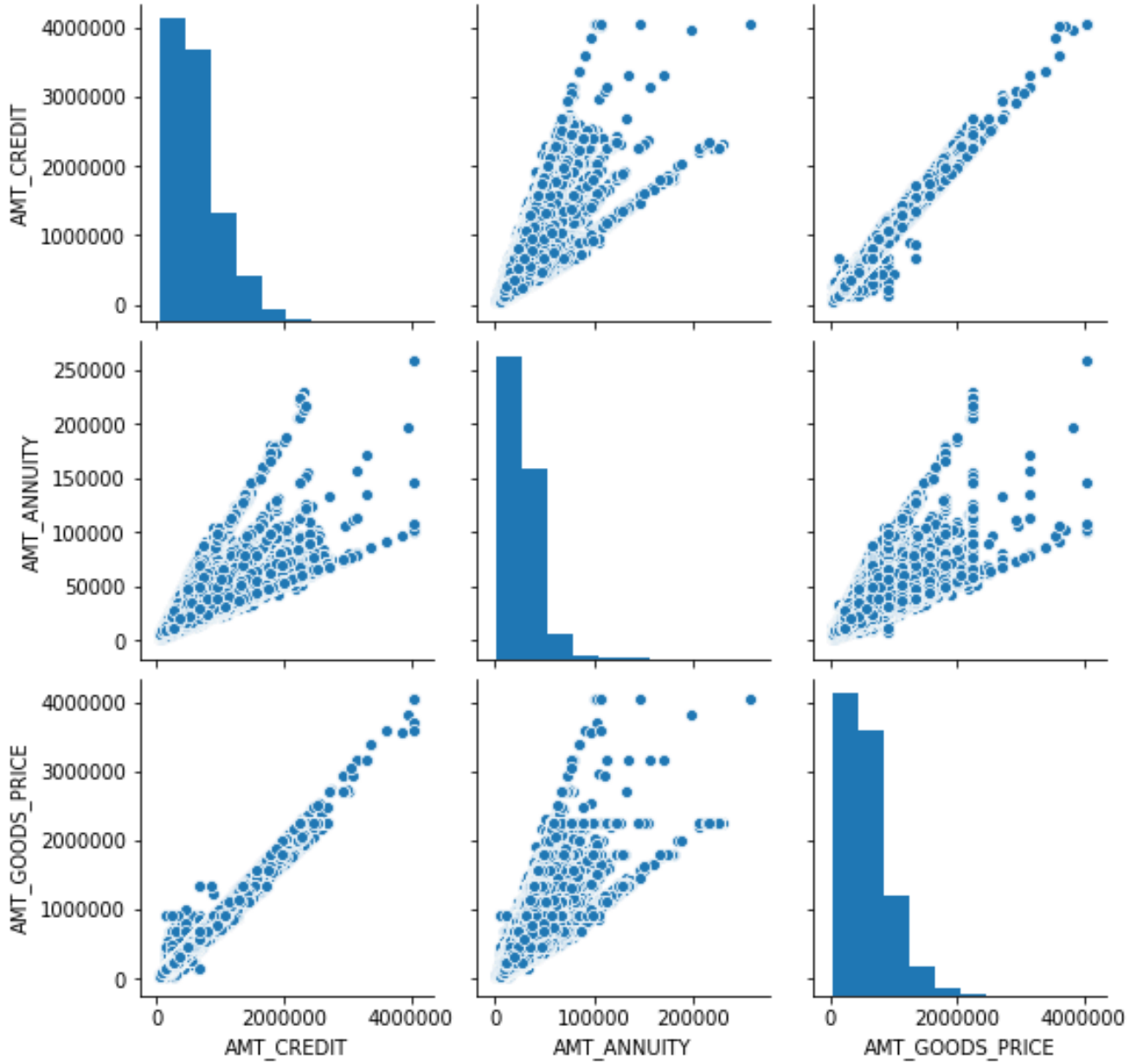
Analyzing Amount of Goods Mortgaged Vs TARGET



Inference:

- We can see an interesting pattern hidden with Annuity Goods and Clients under TARGET = 1.
- While analyzing the percentage that is annotated in the above plot that gives the ratio of clients under TARGET = 1 to Total Clients falling under that particular Goods Amount range, we can see that
- Percentage of Clients under target 1 goes on increasing till Amount of Goods mortgaged up to 5 LPA.
- Post 5 LPA of Goods Amount, the percent of Clients unable to repay loan goes on decreasing.

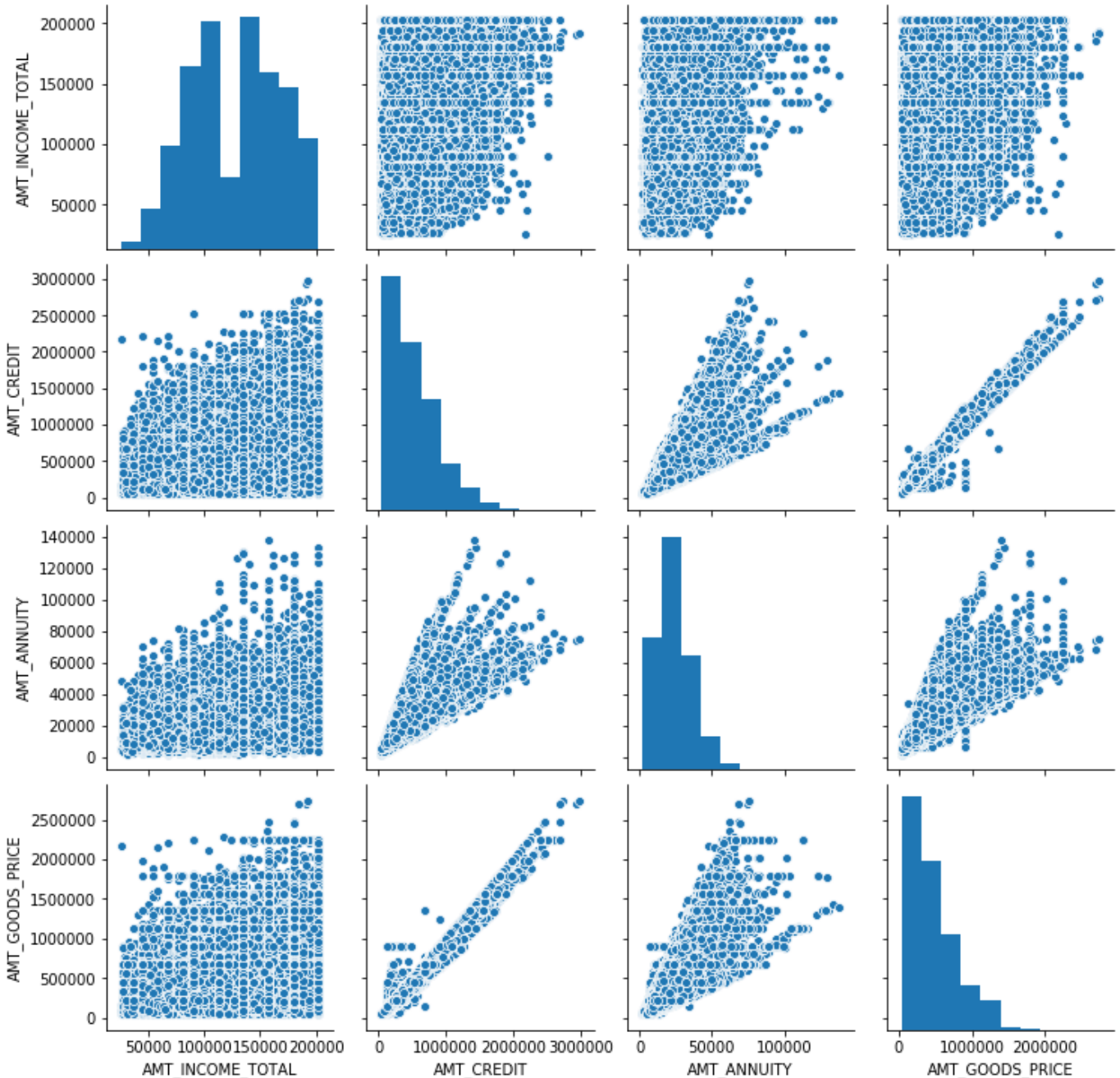
Analyzing relations between Amount Credit, Amount Annuity, Amount Mortgaged



Inference:

- We can see here that all the 3 variables has a positive linear correlation between each other.
- If Credit Amount increases Annuity Amount Increases and Goods Price also Increases and vice versa

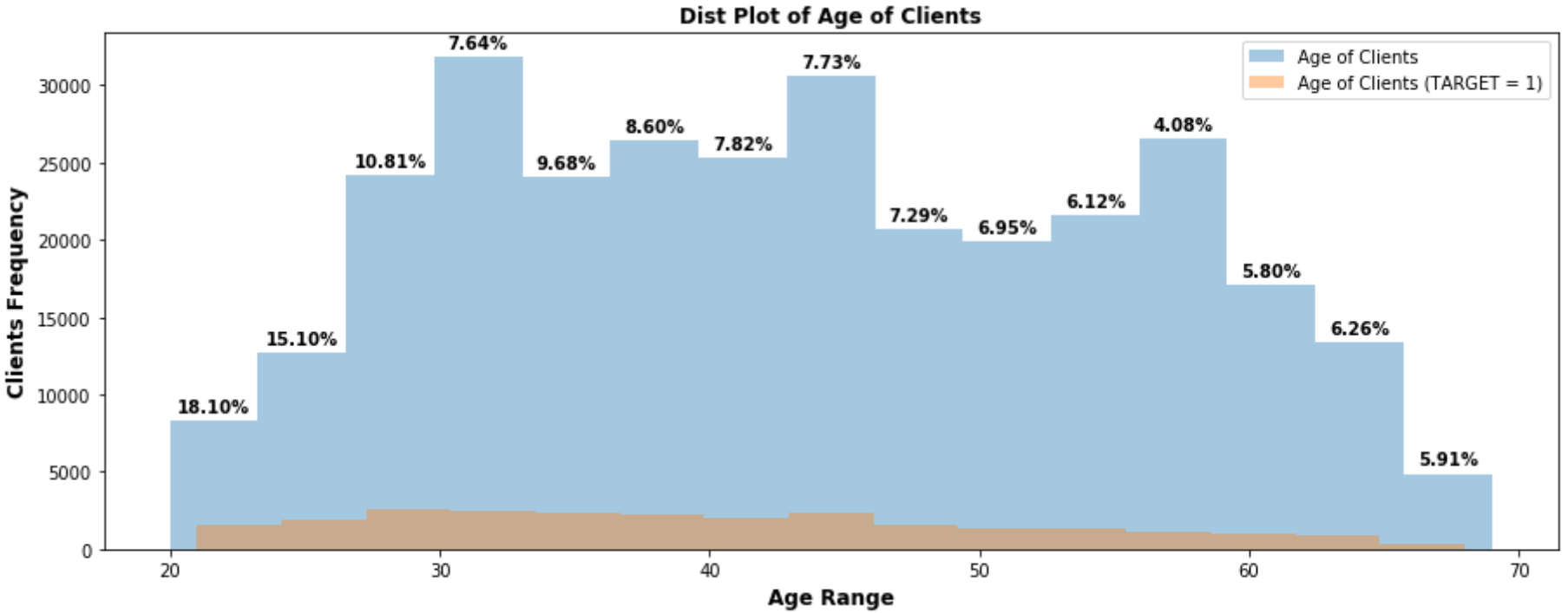
Analyzing relations between Amount Credit, Amount Annuity, Amount Mortgaged and Amount Income



Inference :

- Upon analyzing above graph we can see that there is no strong correlation between Income of Clients with Annuity, Credit and Goods Price

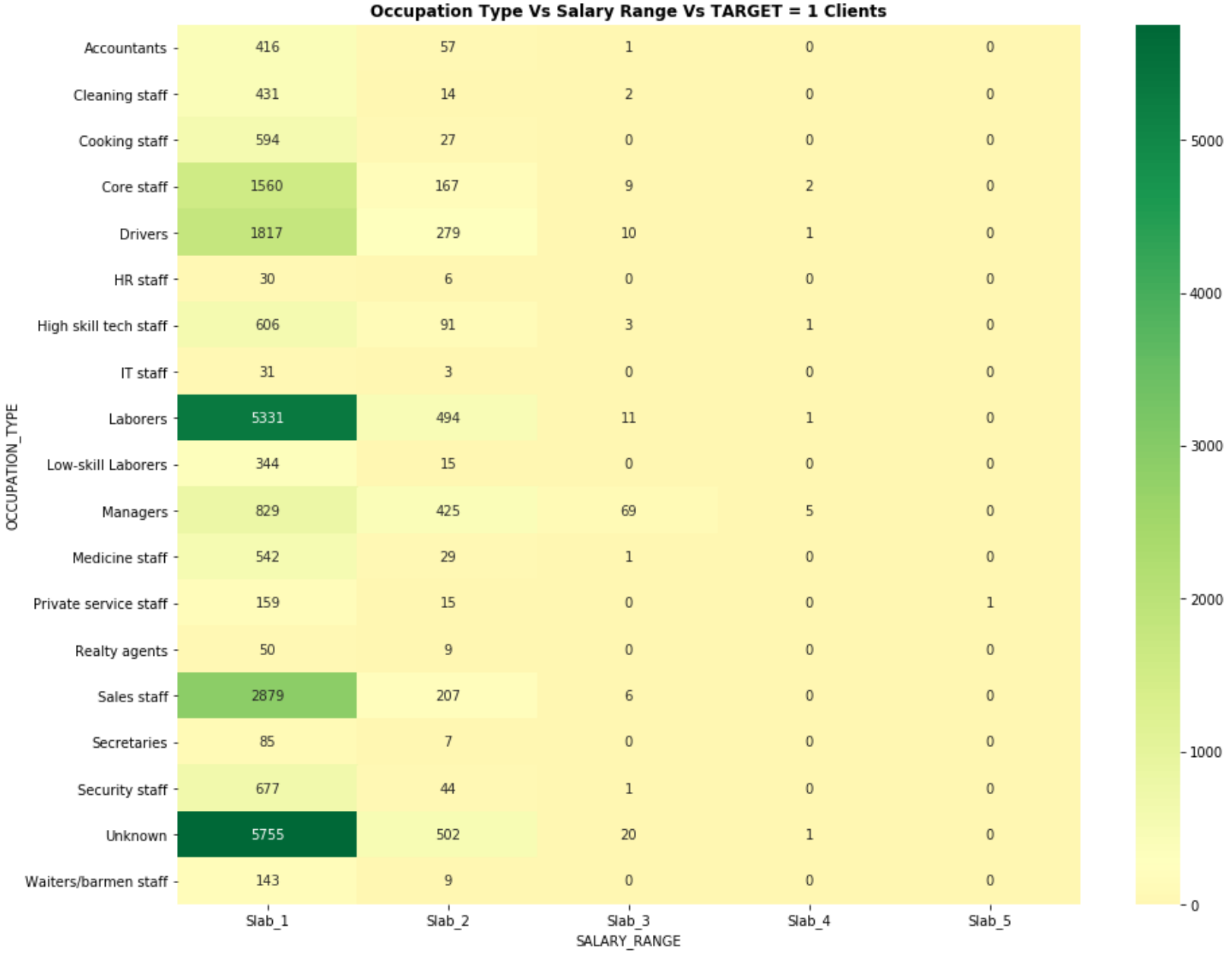
Age Distribution Analysis



Inference :

- Percentage of Clients unable to pay loan is highest in the age range 20 - 23.
- Percentage of Clients according to age group goes on decreasing till 33, And it suddenly increases for age group 34 to 37.
- Then it keep on decreasing gradually.
- It is obvious that banks can focus on age groups from 30 to 33 the most. After which they can focus on age groups from 40 to 60. Also, they can strict the regulations for age groups from 20 to 30

Salary Range Vs Occupation Type

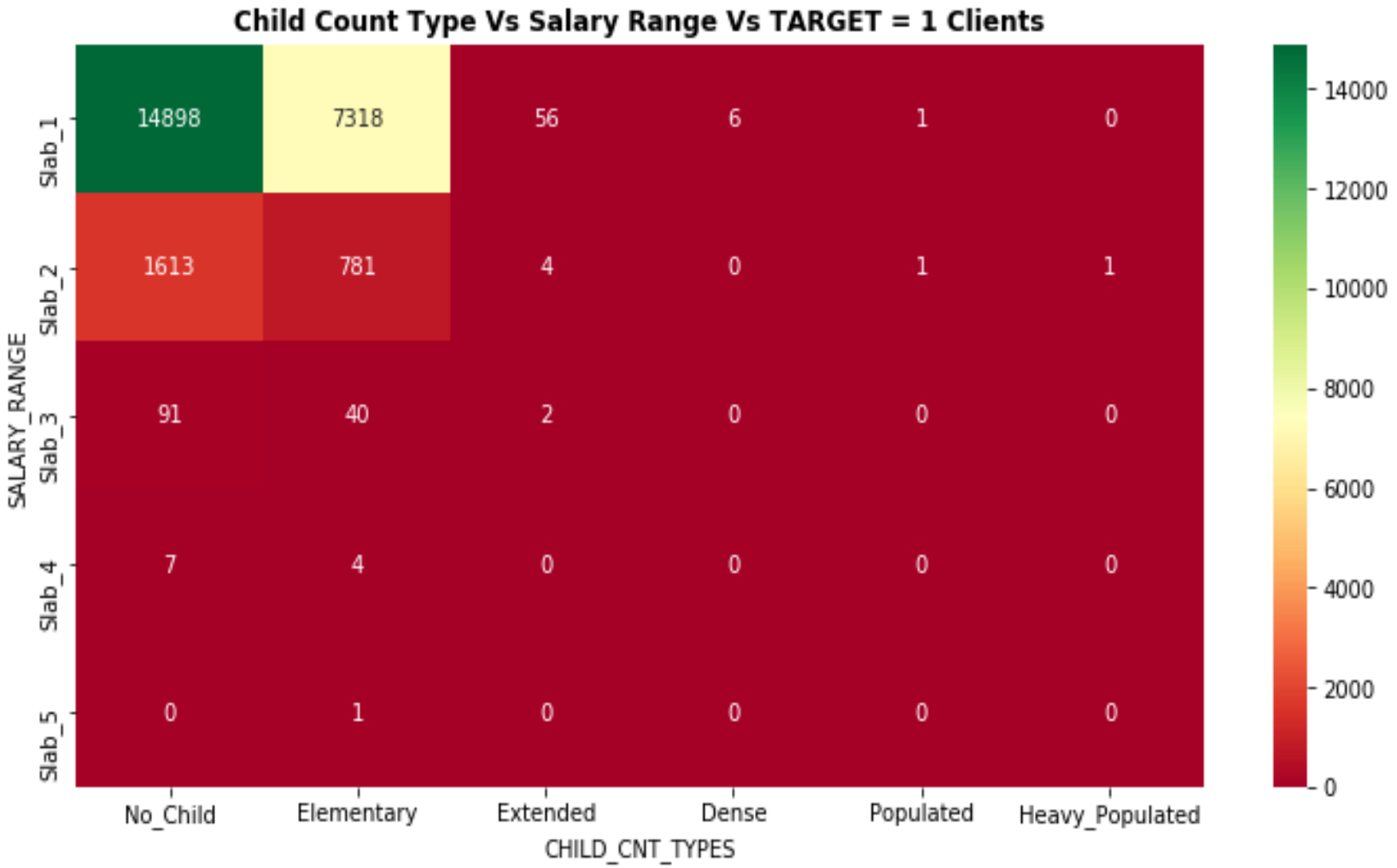


Inference:

- It is obvious that there are considerable number of clients under all Occupation types for slab 1. And it has the most number of defaulters.

- Laborers with salary range Slab 1 has the most number of defaulters.
- Sales Staff under salary range Slab 2 has the second most defaulters count.

Salary Range Vs Number of Children Vs Target

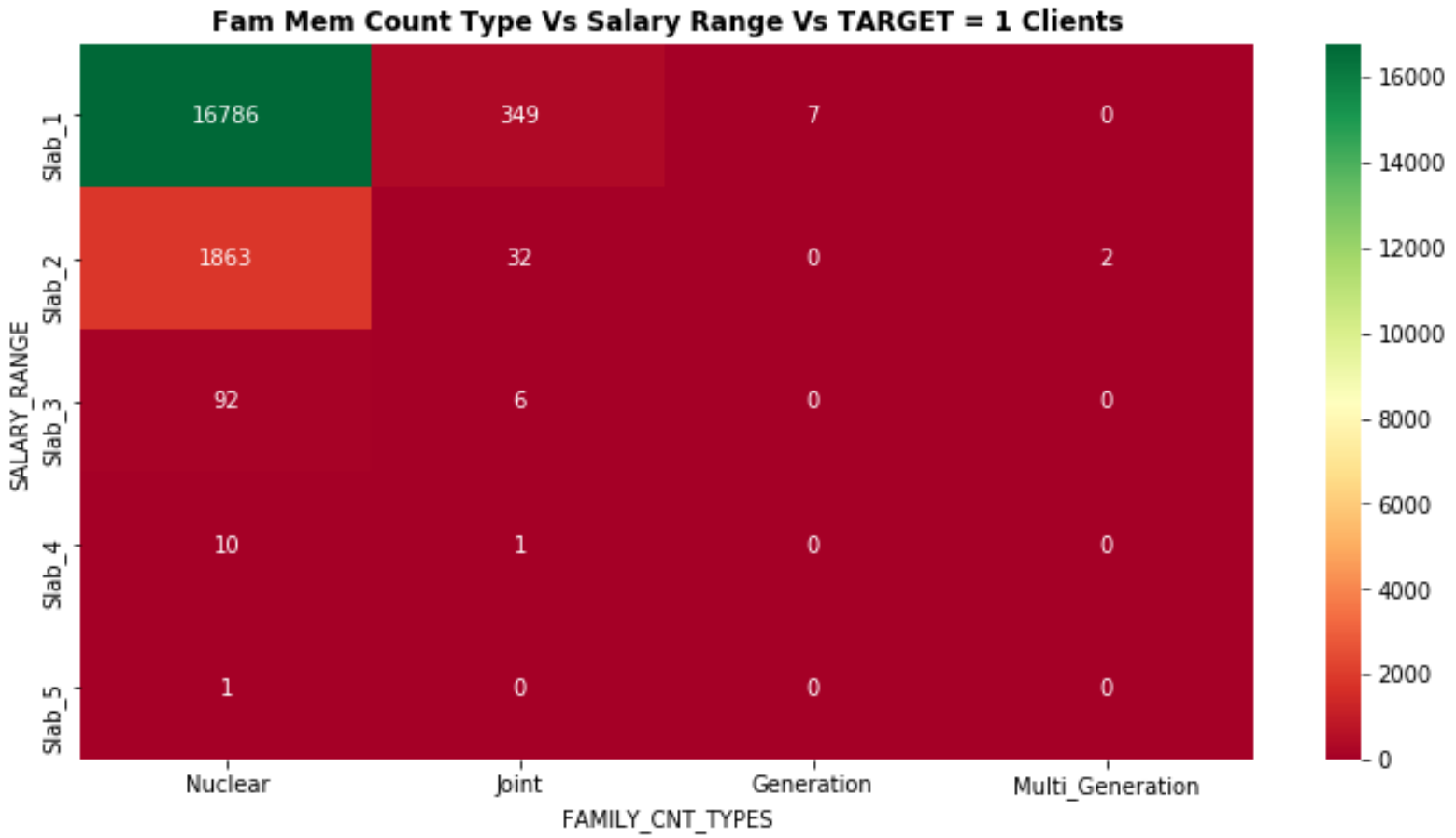


Inference:

Nearly 99% of defaulters falls under only 4 categories

- Slab 1 having No Child
- Slab 1 having 1 to 2 Children (Elementary Family)
- Slab 2 having No Child
- Slab 2 having 1 to 2 Children (Elementary Family)
- Here too, large number of defaulters are under category of Clients falling under Slab 1 having No Child

Salary Range Vs Family Members Count Vs Target = 1



Inference:

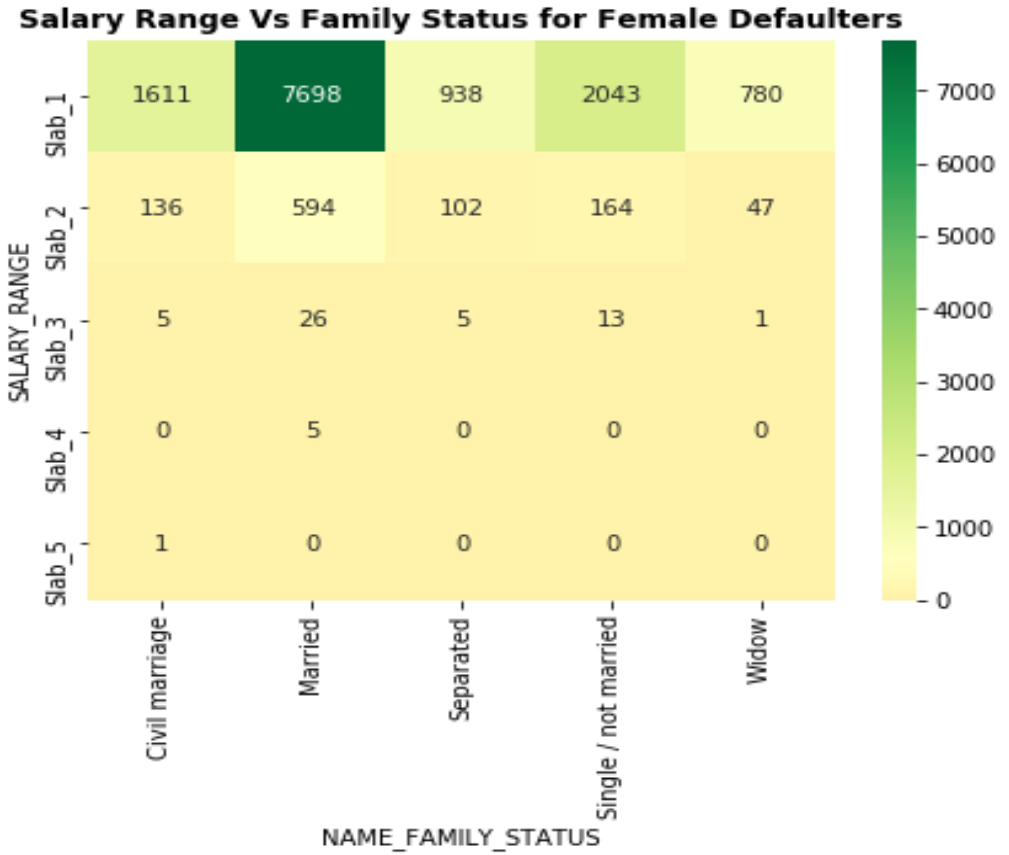
- Here nearly 99% defaulters are under Family type Nuclear (having 1 to 3 family members)

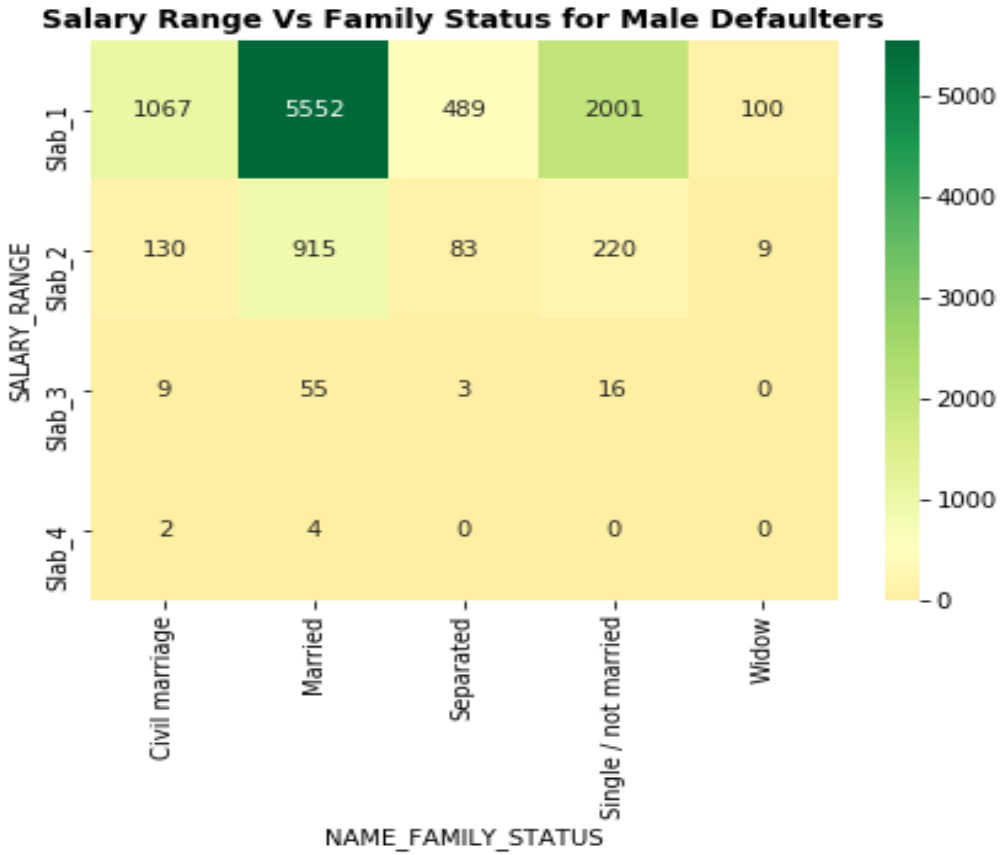
MID SUMMARY:

- So money lending companies has to stringent rules for below category people

- Clients under Slab 1 Salary Range (which is the most important category to focus)
- Clients under Nuclear Family Type (Second most category - Family member count 1 to 3)
- Clients with No Child
- Clients under Occupation type Laborers
- Clients under Occupation Type Business Entity Type 3

Gender Vs Family Status Vs Salary Range (For Target = 1)





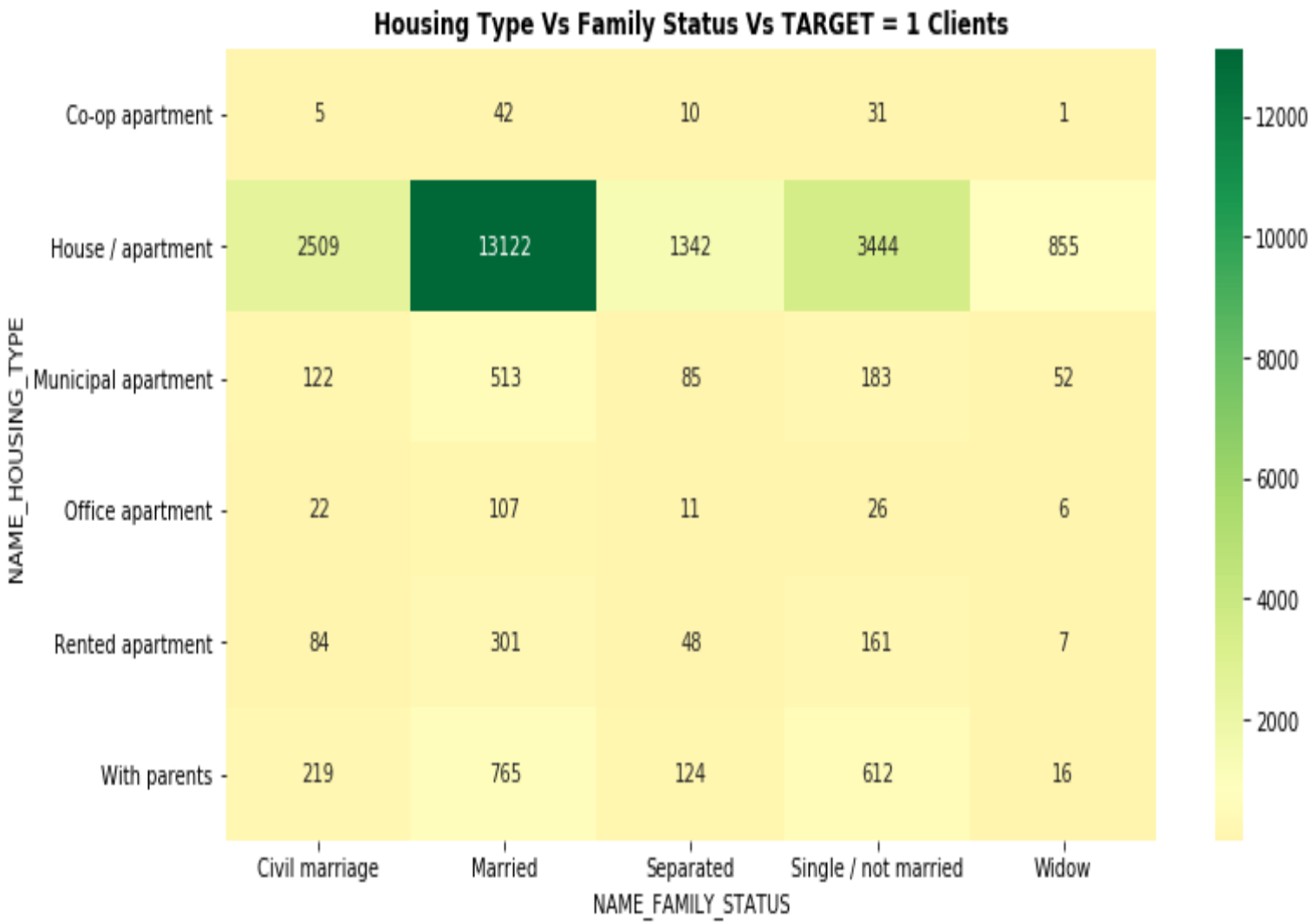
Inference:

- Defaulters are mostly from Slab 1 Salary.
- In both male and Female, most defaulters are from Married and Single / not married Family Status

For Below category people, companies can increase its stringency to provide loans

- Slab 1 Clients
- Clients under category Married
- Secondly, Clients under Category Single / Not Married

Housing Type Vs Family Status Vs Target = 1 Count



Inference:

- The same pattern discussed above is followed here
- Clients with family status Married are the most defaulters.
- Another insight here is Clients with House/Apartment category carries most of defaulters than other group

Occupation Type Vs Education Type Vs Target = 1 Clients

Inference:

- We can infer here that large number of defaulters are under education category Secondary / Secondary Special and Higher Education clients.
- Here too same pattern of defaulters follows with Occupation Type,
- Laborers (The most defaulter)
- Sales Staff (Second most defaulter)

Validating Inferences

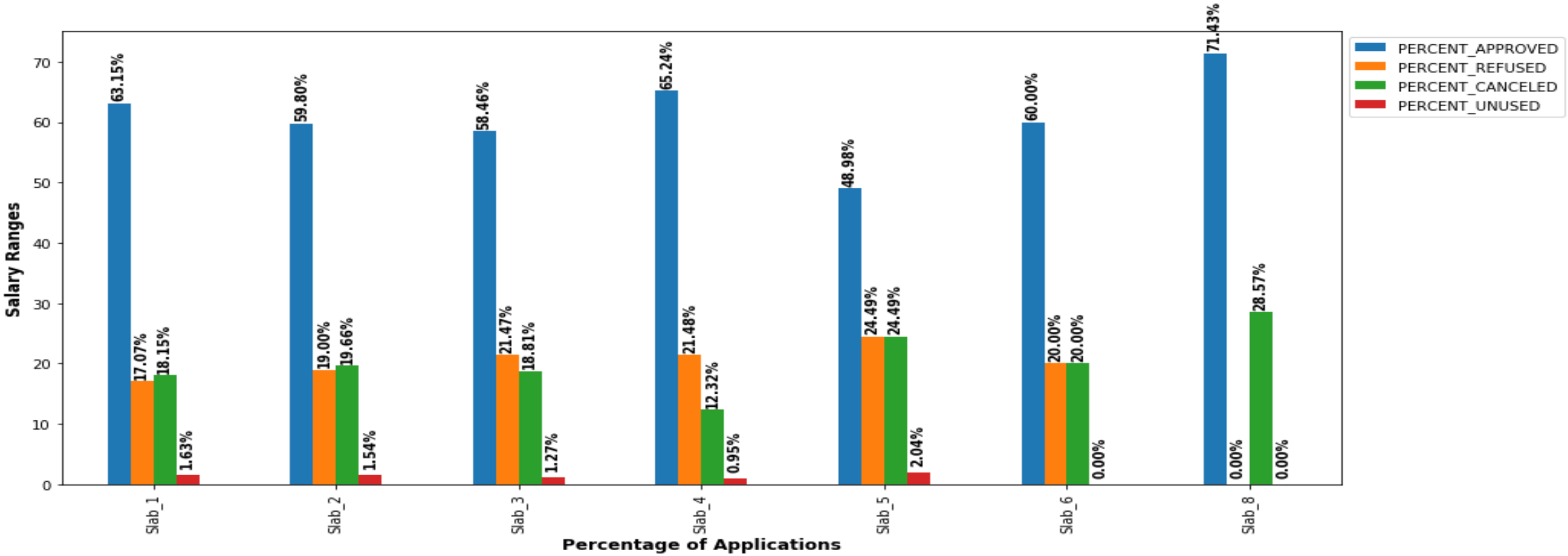
Collecting most contributor categories of defaulters and validating the inference

1. Family Status - Civil Marriage, Single/ Not married
2. Occupation - Unknown, Laborers, Sales Staff
3. Organization - Business Entity type 3, Self Employed, Unknown
4. Applicant's Suite type - Other_B, Other_A, Unaccompanied
5. Income type - Pensioner, Unemployed, Working
6. Education Type - Lower Secondary, Secondary/Secondary Special
7. Housing Type - Rented Apartment, With Parents
8. Income Slab - Slab1, Slab5
9. Family Member Count > 2
10. Children Count <= 1

INFERENCE CONCLUSION :

1. Percentage of Clients becoming Defaulters for above criteria 20.48%
2. Percentage of Clients becoming Defaulters excluding Occupation Type 'Working' for above criteria 50.0%
3. The above verification shows Clients under above categories having 1.05 (approx) out of 5 Clients becomes Defaulters.
4. If Working Clients are excluded, 1 out of 2 Clients are becoming Defaulters

Salary Slab Vs Previous Contract Status



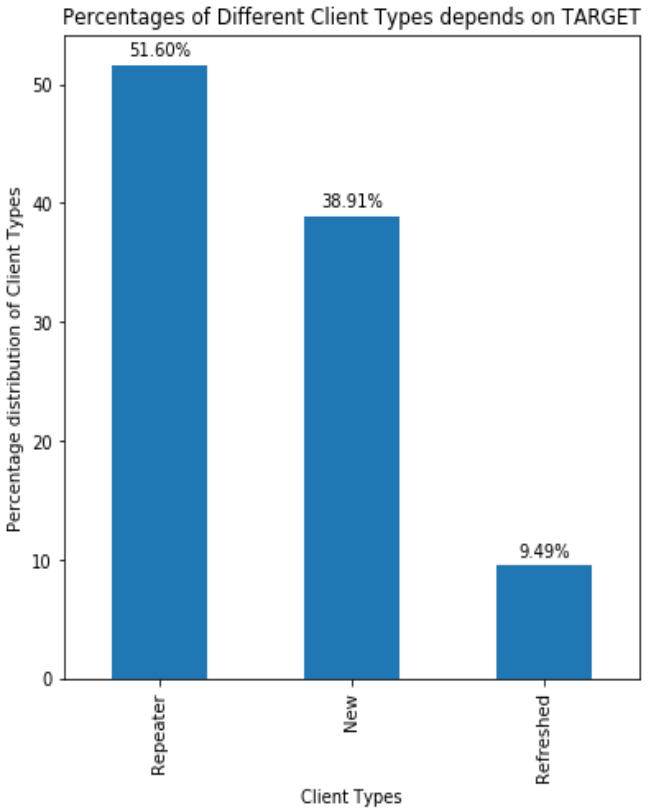
INFERENCE :

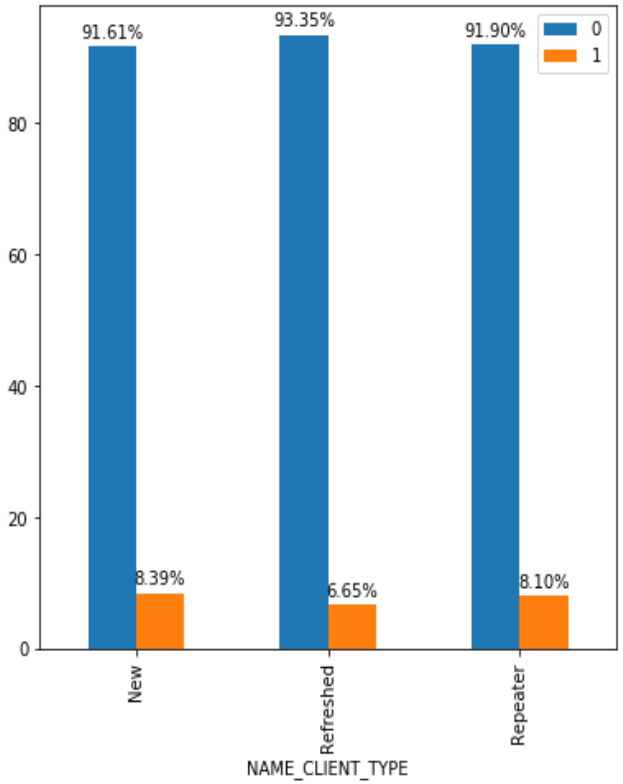
- We can see here that Slab_8 has highest approval percentage of loans.

Approved status of loans follows below pattern regarding Salary Ranges

- Slab 8 > Slab 4 > Slab 1 > Slab 6 > Slab 2 > Slab 3 > Slab 5

Ratio of Client types

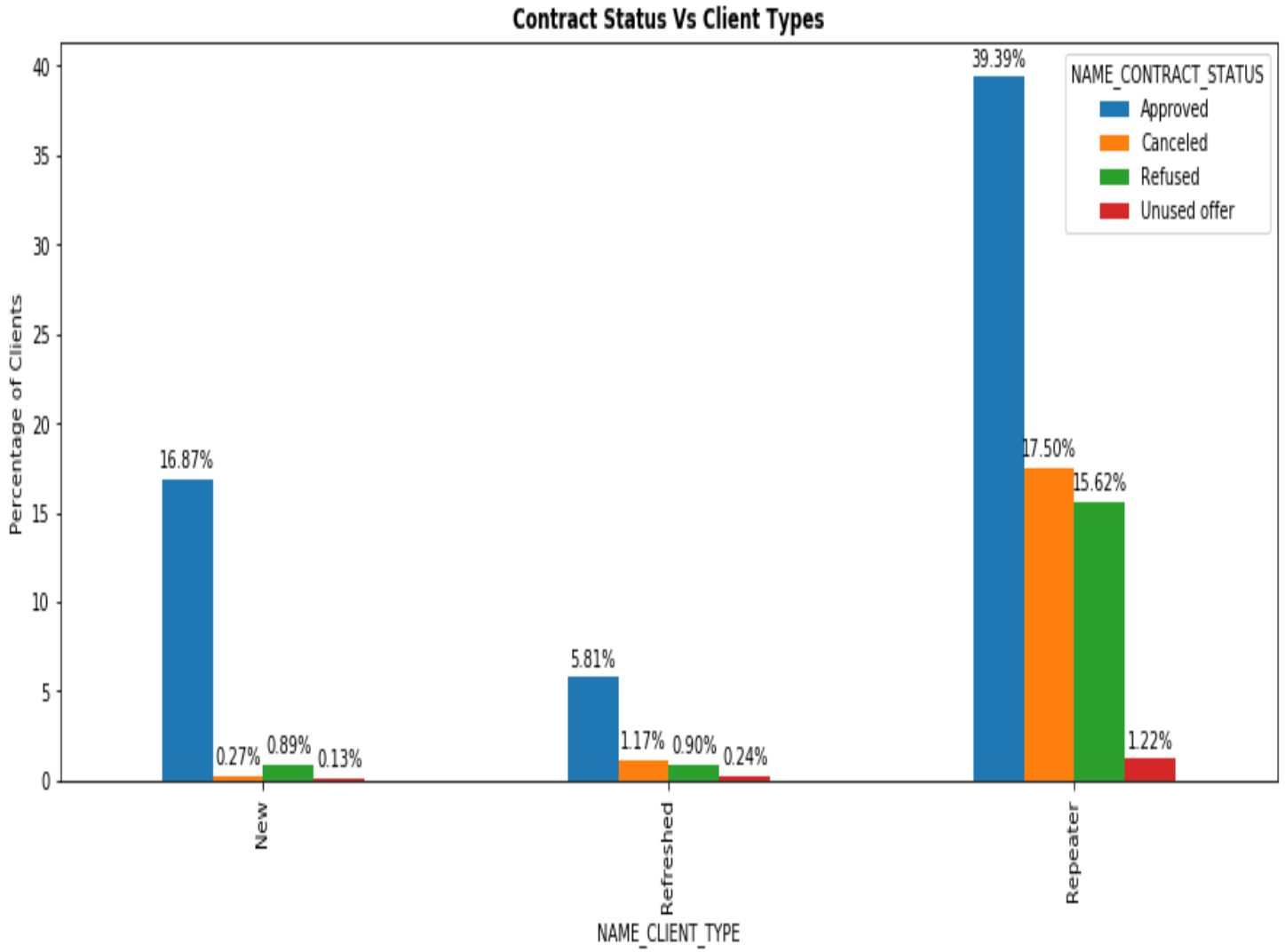




Inference:

- We can see here that nearly 52% of clients in current applications are 'Repeater'. Only 39% of New Applicants applied for the loan
- It is obvious that maximum clients of TARGET = 1, falls under Client Type 'New‘
- So, banks/similar institutions has to take care the most if the client is new for loan application process

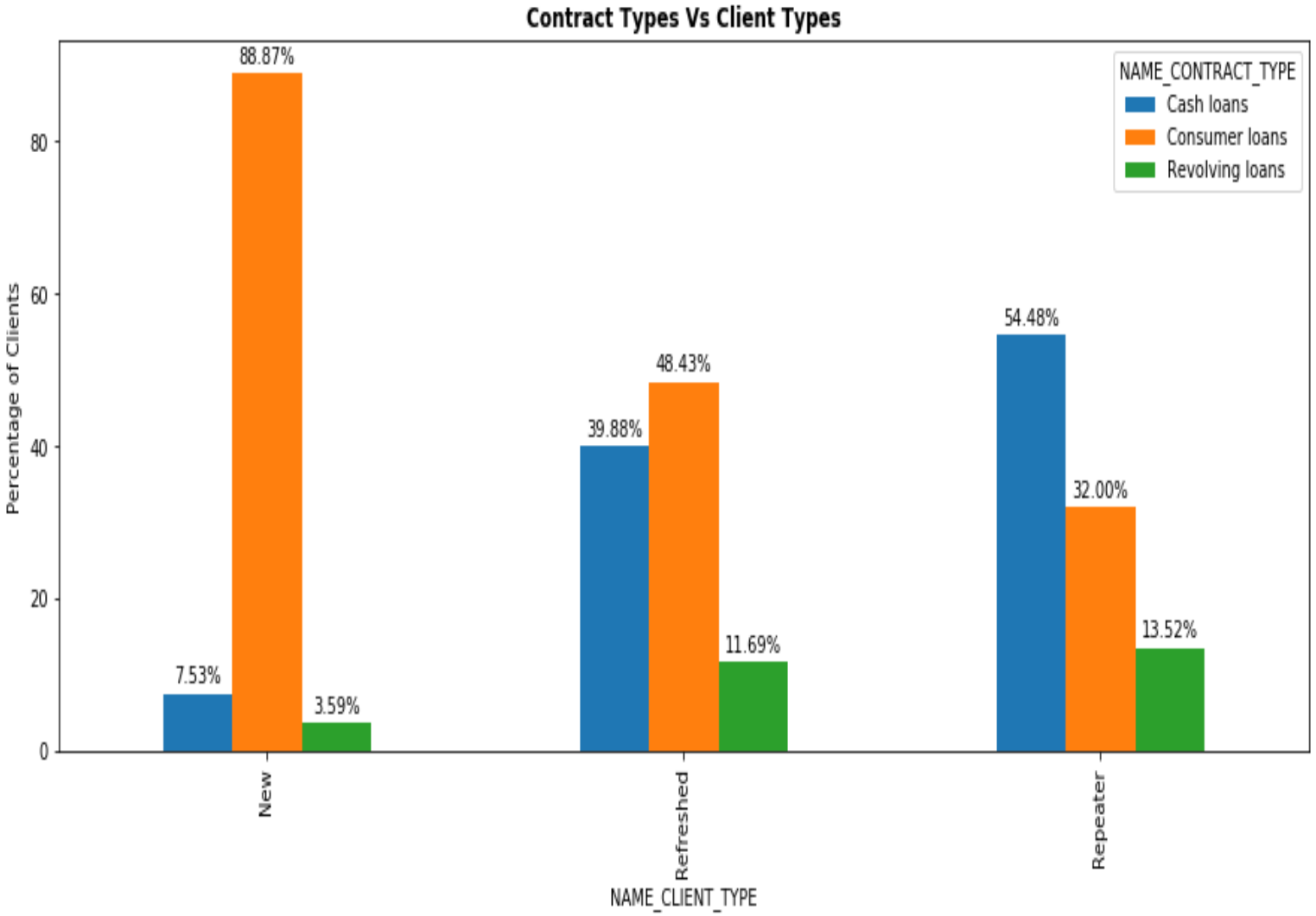
Contract Status Vs Client Types



Inference:

- It is seen that the Approved Status of applications for Repeater Clients are very high than any other client categories.
- The Approved Status follows below Order : Repeater > New > Refreshed
- The Refused status follows below Order : Repeater > Refreshed > New

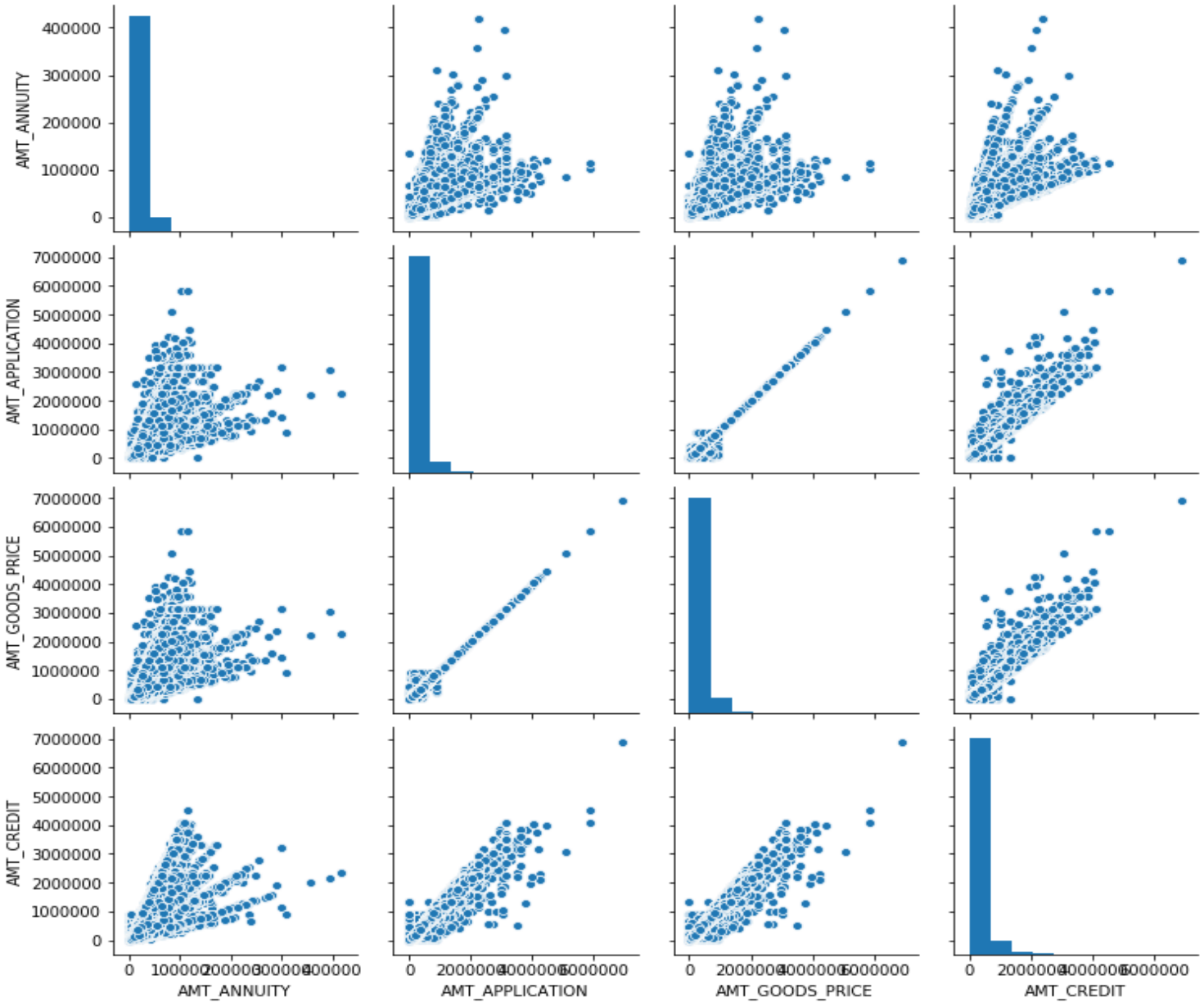
Contract Type Vs Client Types



Inference :

- Clients of Category 'New' has the highest applicants for Consumer loans than the other 2.
- Refreshed and Repeater Clients are having pretty much similar percent of applicants on Revolving Loans.
- Consumer Loans is the highest category of type of loans

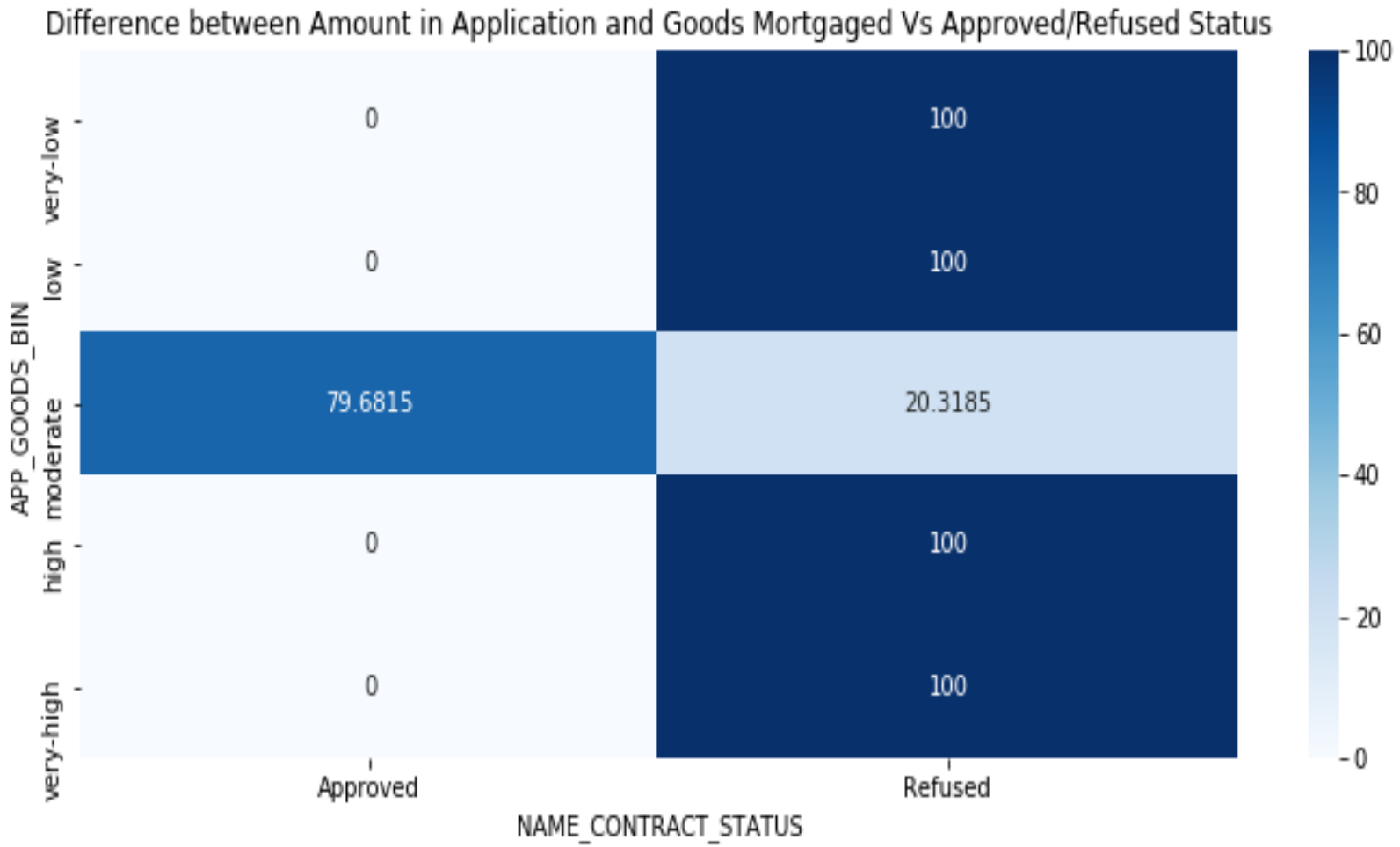
Analysis on Amount Variables



Inference:

- There is a very strong positive relation between Amount of Goods Price and Amount of Application. Also, Goods Price has a strong Positive correlation with Credit Amount.
- Amount Credited and Amount of application also follows a positive correlation

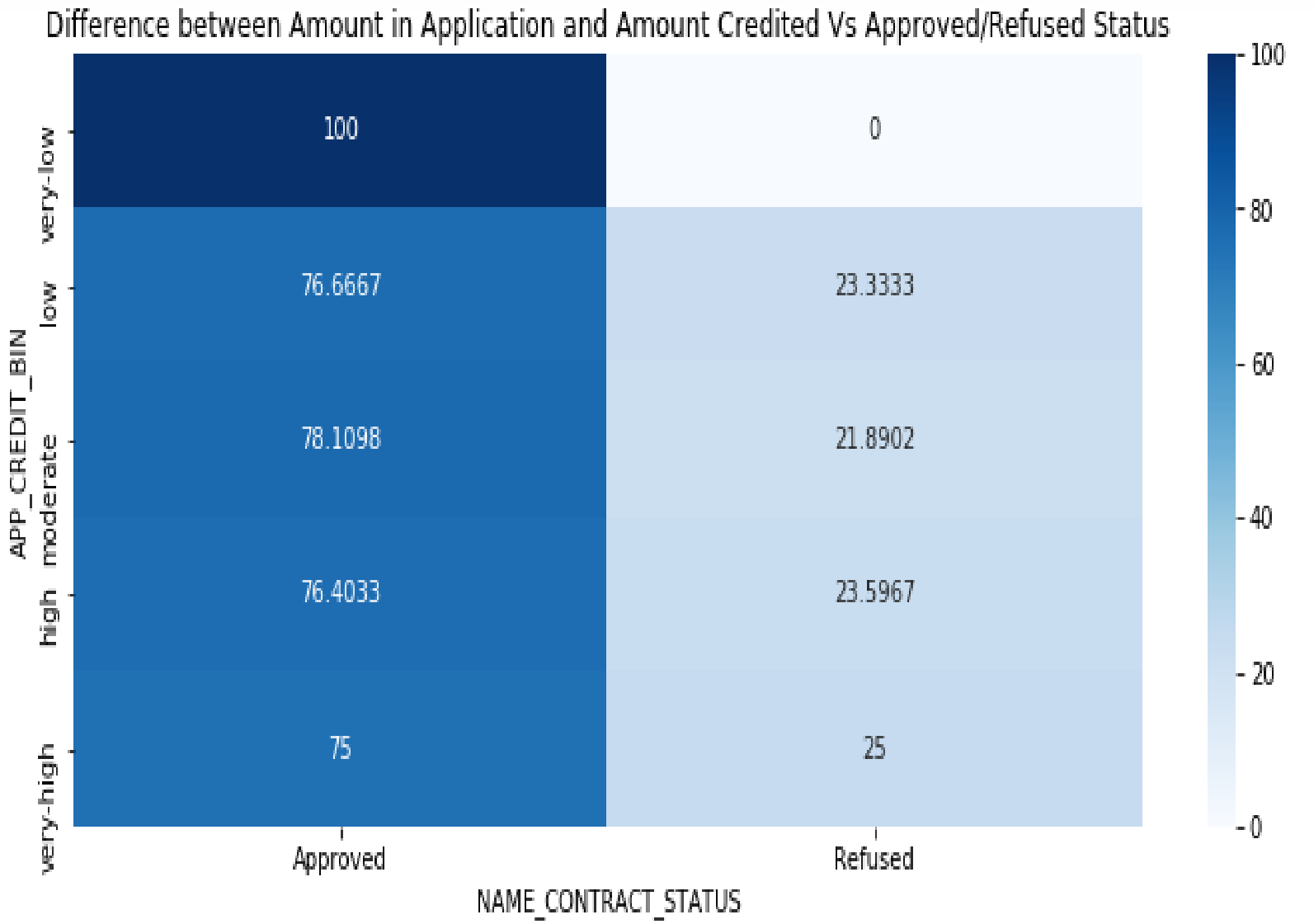
Difference between Application Amount and Goods Mortgaged Amount Vs Approved / Refused Status



Inference:

- It can be visible that only if the applied amount is moderately less than the Mortgaged goods' true price (from -5Lakhs to 0), companies are approving the loans.
- For all the remaining categories, companies are rejecting the loans.

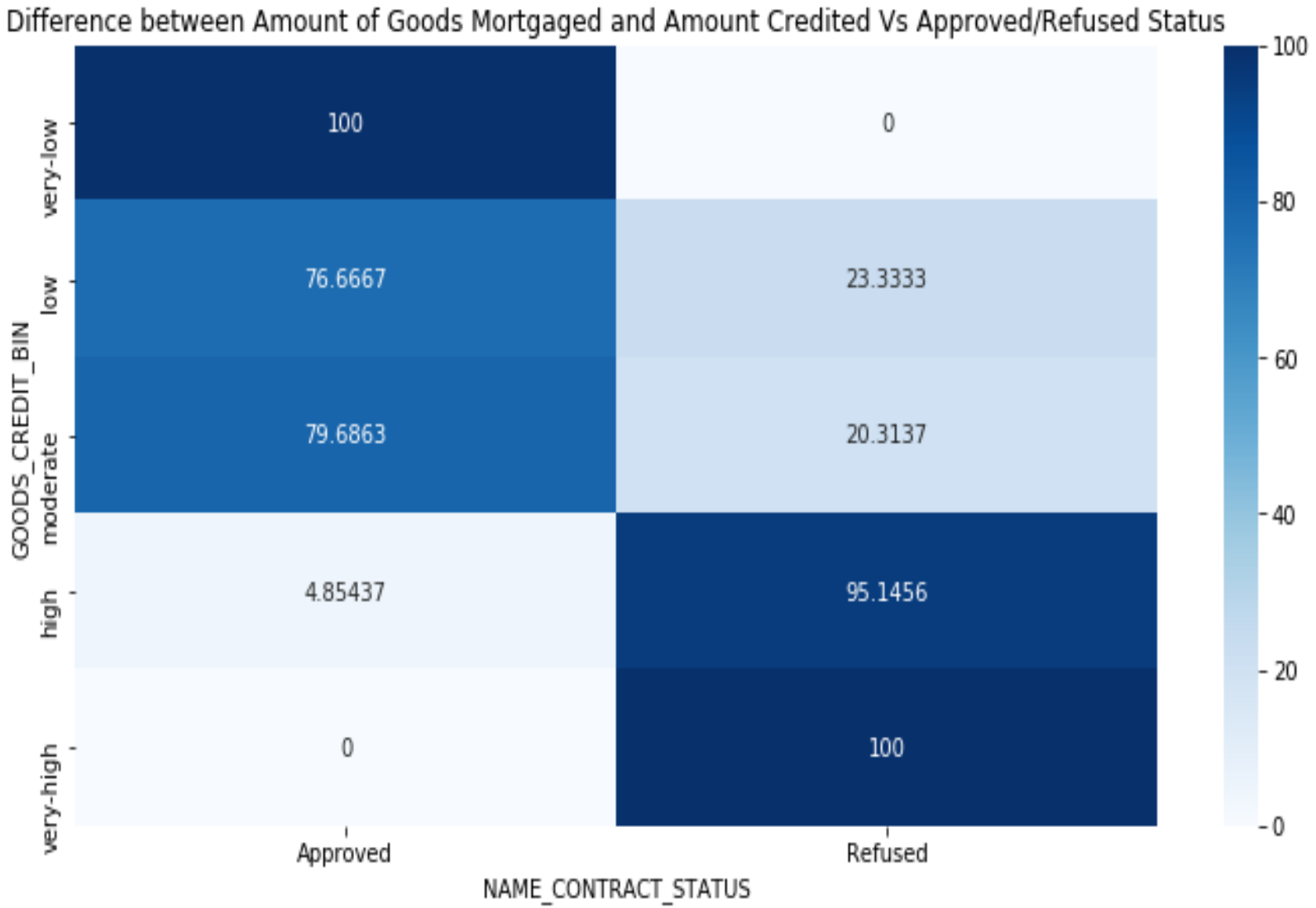
Difference between Application Amount and Credited Amount Vs Approved / Refused Status



Inference :

- It is interesting to see that there are loans where the companies credited the amount greater than the amount of application.
- It is seen that for those loans where the loan amount credited is lower than the loan amount applied (nearly -35Lakhs to -20Lakhs - which is the lowest category among the bins), there are almost flawless approval of loans i.e., near to 100%
- If the difference between the loan amount credited and loan amount applied is increases, then the companies tend to approve only 75%(approx) of total loans.

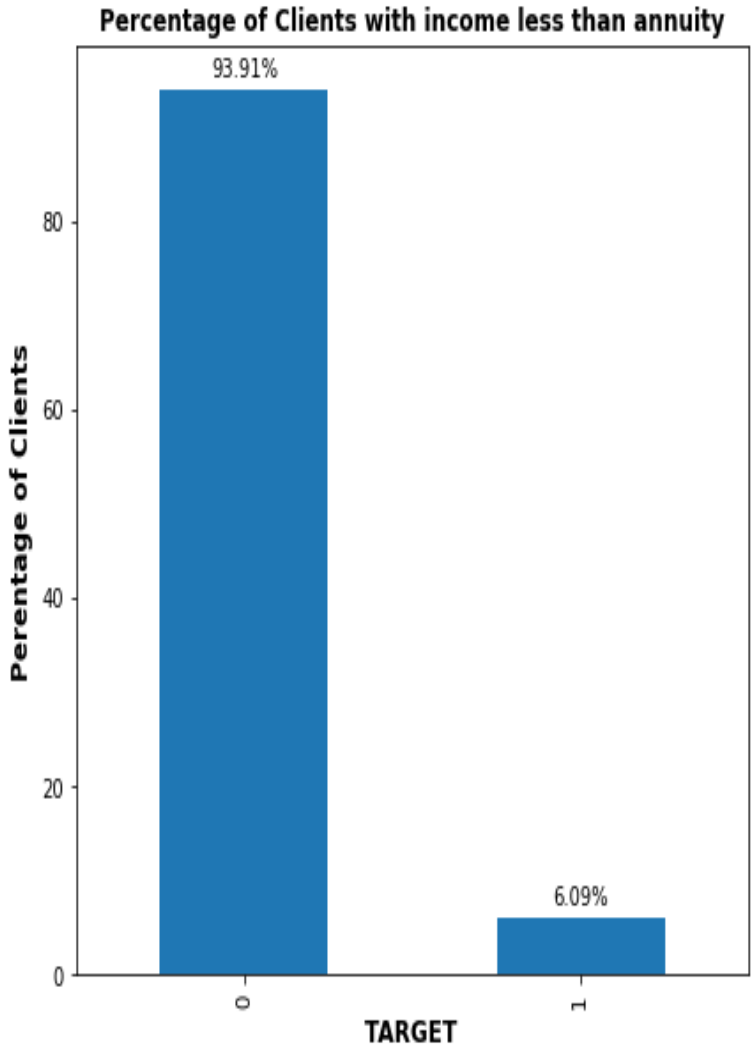
Difference between Goods Mortgaged Amount and Credited Amount Vs Approved / Refused Status

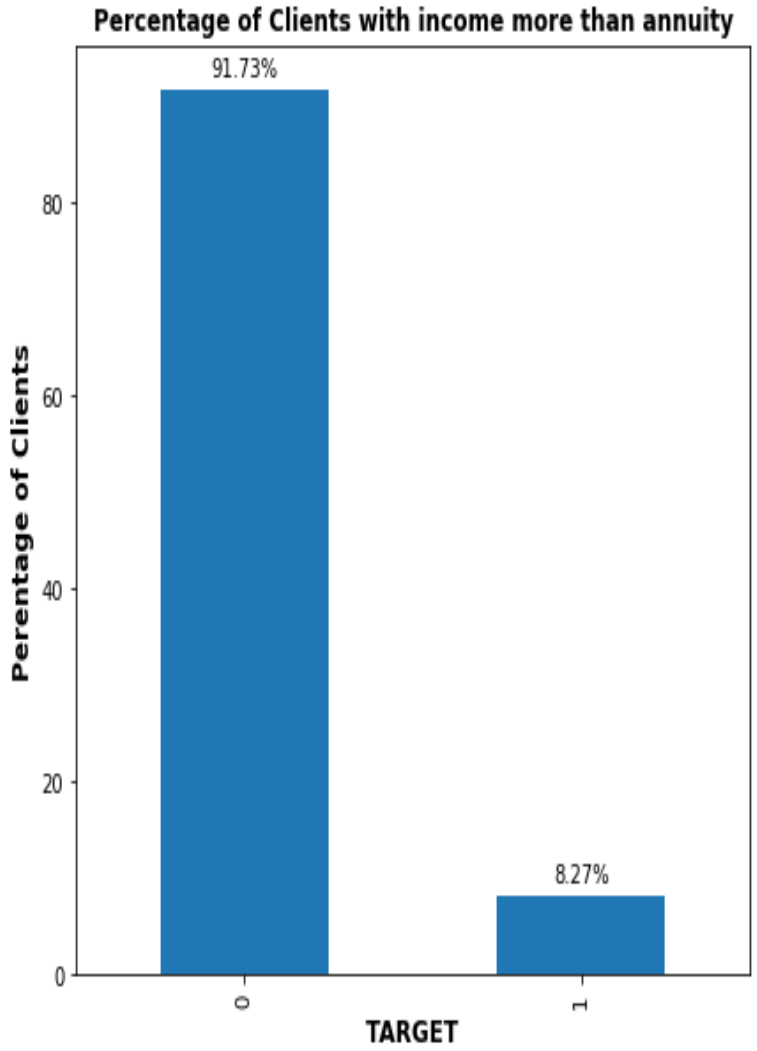


Inference:

- We can get an interesting insight with above plot that there is no or meagre number of loans approved (4.8%) with amount credited value is greater than the amount of goods mortgaged.
- If the amount credited is very much less than true amount of goods mortgaged, then there are high number of approved loans and it reaches to 100% if the difference is quite low.

Analyzing relation between Approved and Cancelled loans with difference in Amount of income of Clients with Amount Annuity

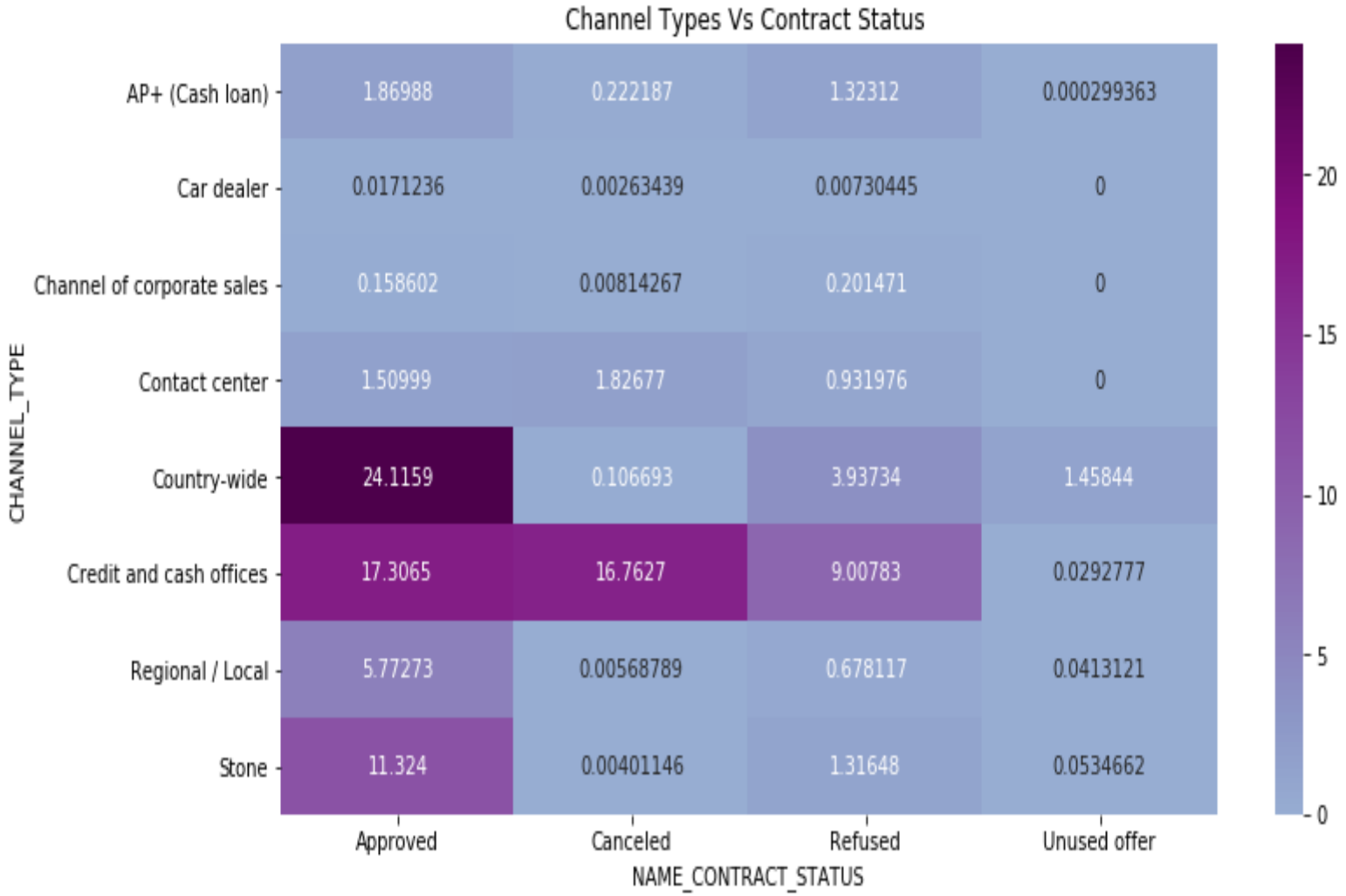




Inference :

- This plot is showing that Clients with income higher than annuity is falling under defaulter category than the Clients with income less than annuity.
- This can be due to several other factors like owning house or not, number of children, family members etc

Finding the successful Chanel Type



Inference:

- The most familiar channel type for Approved loans is 'Country-wide' , 'Credit and cash loans', 'Stone‘
- 'Credit and cash loans' have equal opportunities that a loan can be approved by Company or Canceled by Clients
- The least familiar channel type is 'Car Dealer'

Conclusion

With our analysis on the application_data.csv, we inferred below insights,

- TARGET is the variable based on which we have to analyze. It has higher ratio of Data Imbalance. It is a categorical variable with only 2 categories (1 - for Defaulter Clients ; 0 - Clients Paying the loan properly)
- The imbalance ratio is about 92% for TARGET = 0 and 8% for TARGET = 1.
- Females are the highest applicants of loan. About 66% of Applicants. But for TARGET = 1 , Female defaulters are somewhat less (only 57%) and males are having higher ratio of defaulters (43%)
- Percentage of Clients unable to pay loan is approximately closer (9.9% approx.) for Clients in Family status Civil Marriage and Single / not married category
- Laborers, Sales Staff, Core Staff, Managers and Drivers are those Occupation Types where the defaulters are the most
- Business Entity Type 3, Self - Employed, Others, Business Entity Type 2 and Construction are the Organization types where the defaulters are the most.
- Irrespective of Clients owning Flat or not, the defaulters ratio is almost 8% of Clients owning realty or not
- 8% of clients not owning car and 7% of clients owning car are falling under defaulters.
- There is no huge difference in percentage of people who are not repaying loans with Accompany status as unaccompanied with other Accompany types
- Banks/similar companies have to strict the rules to approve loans for Clients in Income types Unemployed, Student, Businessman, Maternity leave

- The category pattern with respect to education Type who are unable to repay loans follows the below format:
- Lower Secondary > Secondary / Secondary Special > Incomplete Higher > Higher education > Academic Degree
- Rented Apartments and with Parents - Housing types which has higher percentage of defaulters.
- It is obvious that banks can focus on age groups from 30 to 33 the most. After which they can focus on age groups from 40 to 60. Also, they can strict the regulations for age groups from 20 to 30
- With respect to Organization Type, Business Entity Type 3 and Self Employed have high number of defaulters. With respect to Income type, Working and Pensioner have high number of defaulters.
- Laborers with salary range Slab 1 has the most number of defaulters. Sales Staff under salary range Slab 2 has the second most defaulters count.
- Large number of defaulters are under category of Clients falling under Slab 1 having No Child
- Money lending companies has to stringent rules for below category people
    - Clients under Slab 1 Salary Range (which is the most important category to focus)
    - Clients under Nuclear Family Type (Second most category - Family member count 1 to 3)
    - Clients with No Child
    - Clients under Occupation type Laborers
    - Clients under Occupation Type Business Entity Type 3
- For Below category people, companies can increase its stringency to provide loans
    - Slab 1 Clients
    - Clients under category Married
    - Secondly, Clients under Category Single / Not Married
    - Large number of defaulters are under education category Secondary / Secondary Special and Higher Education clients.