Bank Loan Case Study :
Final Project-2

By: Arpan Sharma(Data Analytics Trainee)

Project Description:-

Bank Loan Case Study is about finding trends and insights about loan application dataset from the bank. In this project, I have used the application_data and previous_application dataset provided by trainity and drawn some conclusions. **The aim of this project is to identify the customers who are capable of repaying loans.** I have provided insights to topics and answered the questions asked by the management team. I have used Google Collab and Google sheets for data analytics and data visualization. **In this case study, I will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.**

Approach:-

Firstly, I have used the basics of the data analytics process to clean the raw data and ask questions from cleaned data. Then, I have used data wrangling to make small data frames for relevant insights to answer all the possible questions. I have used python programming language to write the procedure, functions and creating charts/plots. Finally, I combined all the results and visuals into this report.

Tech-Stack Used:-

I have used the web based application **"Google collab"** which is an online python notebook for performing data analytics using python programming language and "**Google Sheets**" which is part of google online docs for performing various functions on spreadsheets. Both of these software provide ease of work and make data sharing and real time tracking very easy.

Project Insights:-

Table Details:

Table Name	No. of Rows	No. of Columns	Description
application_data	307511	122	122 attributes of bank loan applicant related data of 307511 applicants
previous_applications	1670214	37	37 attributes of bank loan applicant related data of 1670214 applicants
columns_description	160	4	Description of all attributes of both application_data and previous_application dataset

Python Notebook link:	https://colab.research.google.com/drive/1YygKU8qvfjay2kKLWroQ3fXfNxRHe_0i?usp=sharing

A. Cleaning the Data:-

Removing Duplicates: Duplicated values are not present in any of the dataset tables.

Handling Null values:

In application_data,

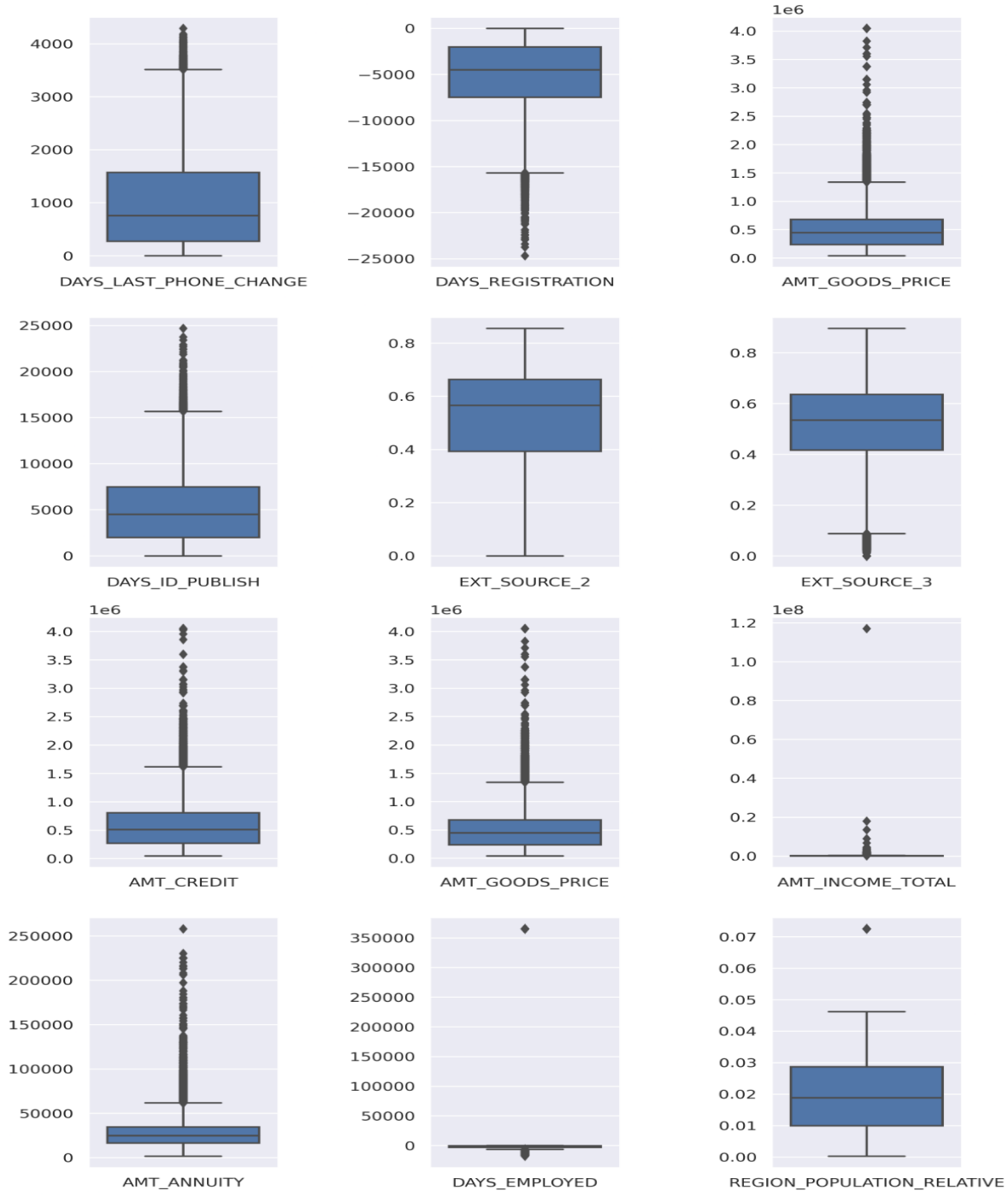
- Out of 122 columns, 49 columns are removed having more than 40% null values and 73 columns are left.
- OCCUPATION_TYPE is important variable with 31.345545% missing data. So, I will replace the null values with the "Unknown" keyword.
- To replace null values in numerical data type columns i.e., columns with int and float datatype, I will use the Median value of the column to replace them as the data may have outliers.
- To replace null values in the NAME_TYPE_SUIT column, I will use the MODE value of the column to replace them as the column has only 0.42% NULL Values.

In previous_application,

- Out of 37 columns, 11 columns were removed having more than 40% null values and only 26 columns left.
- Only 5 Columns have null values. Out of 5, 'AMT_GOODS_PRICE', 'AMT_ANNUITY',' CNT_PAYMENT', 'AMT_CREDIT' are numerical columns and 'PRODUCT_COMBINATION' is categorical variable. I am using Median value to replace null values in numerical columns and mode value to replace those in categorical columns.

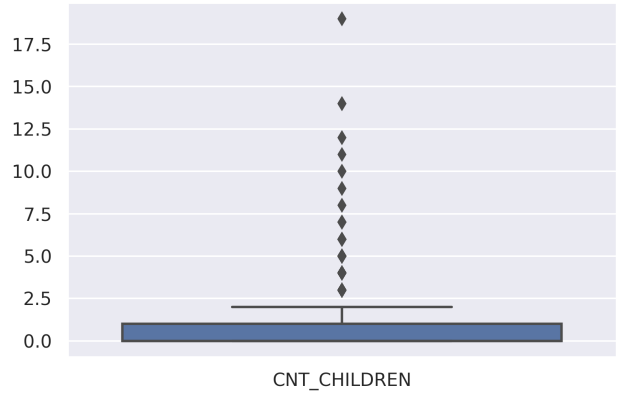
Finding Outliers:

Outliers in application_data dataset:-

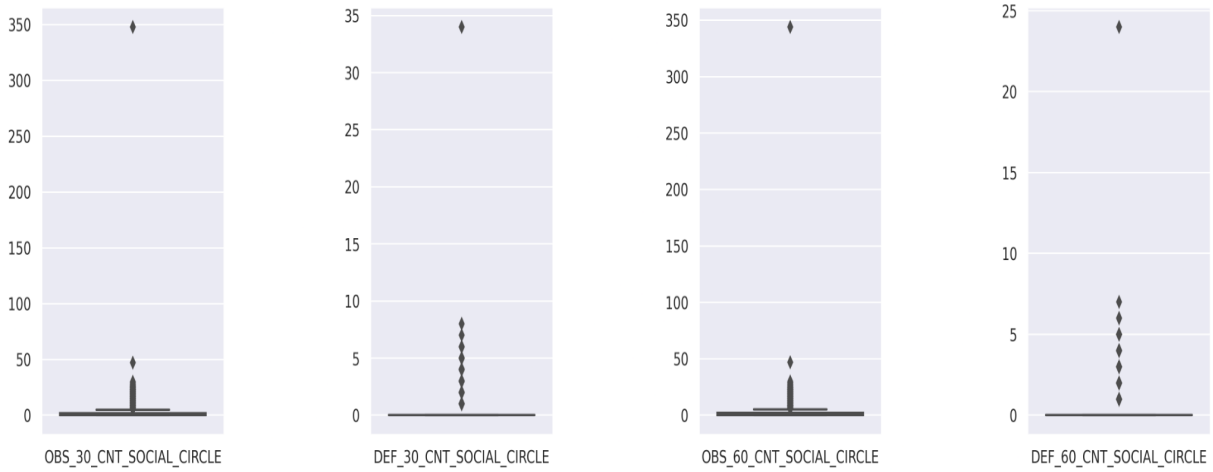


Key takeaway: These box plots of some continuous variables from application_data represent outliers on the basis of quartile range. Out of 12, 11 attributes have outliers but only 4 have extreme outliers. These 4 include AMT_INCOME_TOTAL, AMT_ANNUITY, DAYS_EMPLOYED, REGION_POPULATION_RELATIVE.

- In 'AMT_INCOME_TOTAL', this is an outlier with extremely high income that is around 11.7 crores which is not possible for the occupation type labor working in Organization Type -'Business Entity Type 3'. This record doesn't seem to have a correct value of the income as a labor's income is usually not so high. We can impute this incorrect value with the average income of Laborers working in Organization Type -'Business Entity Type 3'.
- The outlier in this attribute is days employed before application date is more than 350000 days. Which is not possible as it corresponds to nearly 1000 years of employment. These outliers are records of applicants who are either pensioners or unemployed. I will impute this value with 0.

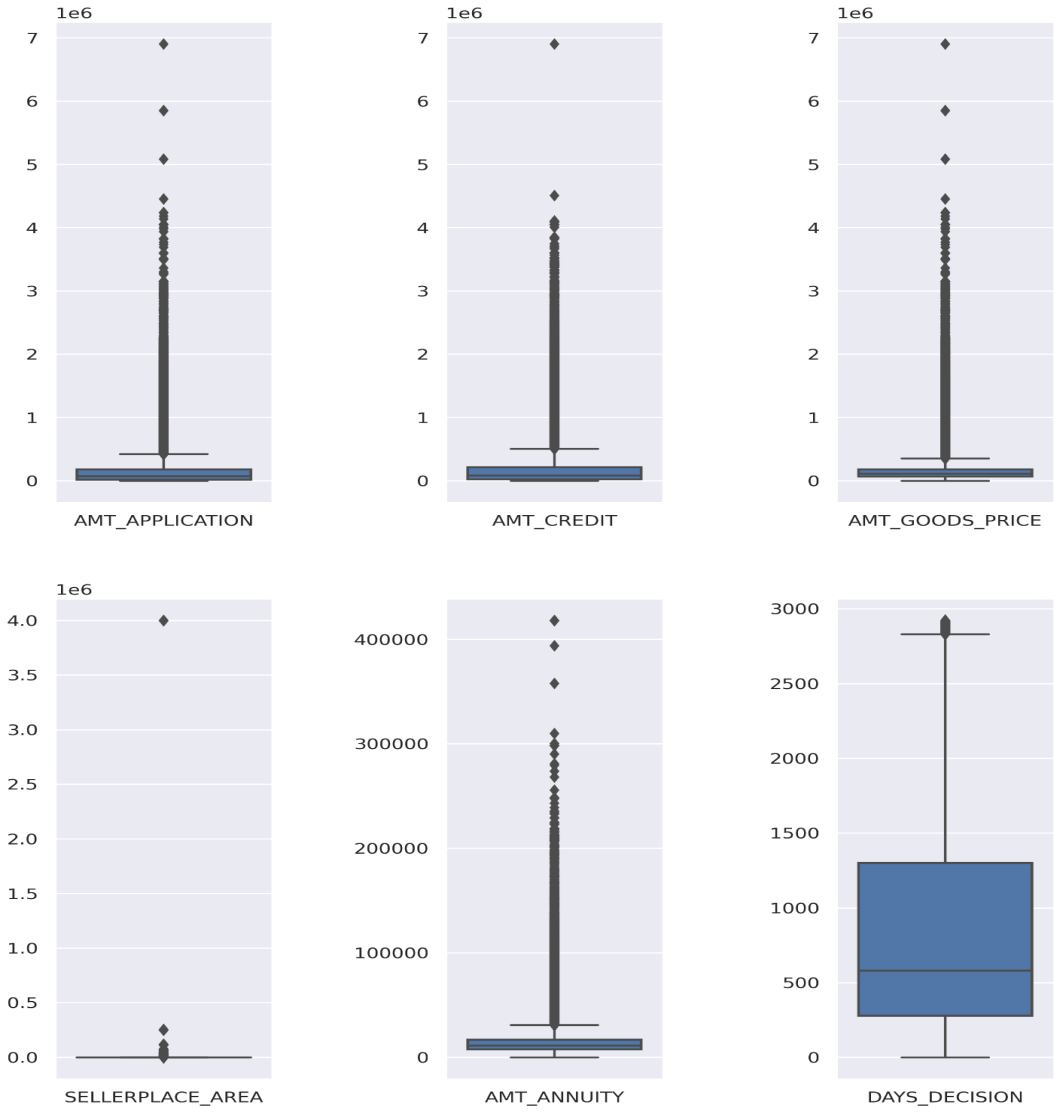


- Two applicants at row 155369 and 265784 are female of age 30 and 28 years and have 19 children which is an anomaly in data.



- Row number 148403 is an outlier in all 4 social surrounding observation for default columns.

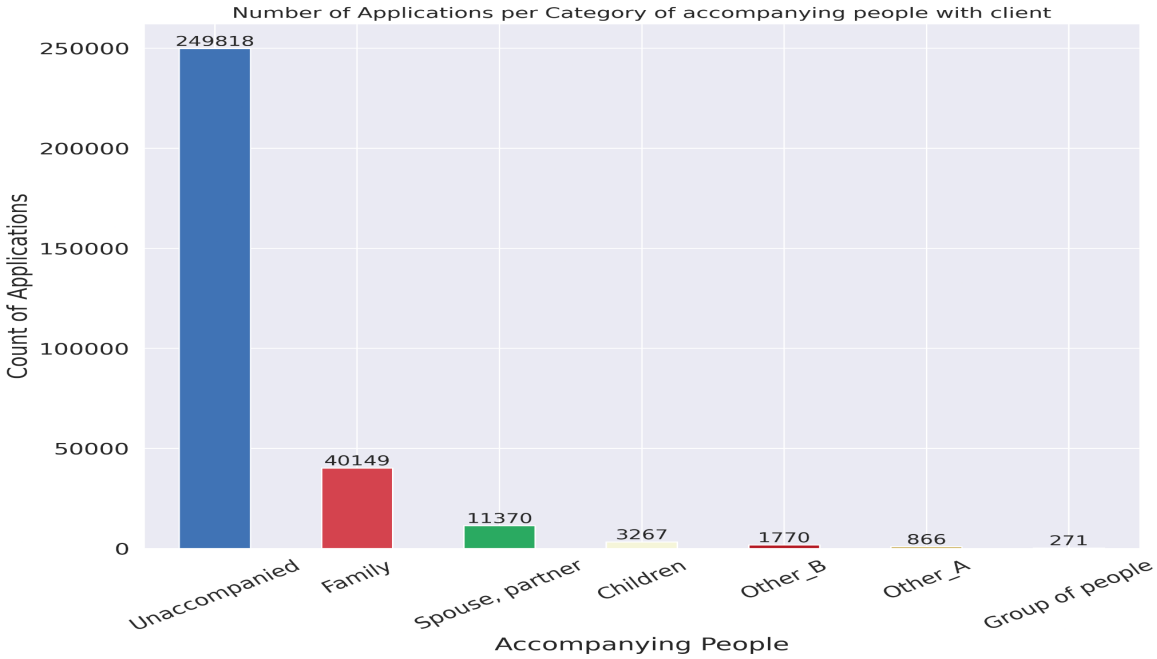
Outliers in Previous_application dataset:-



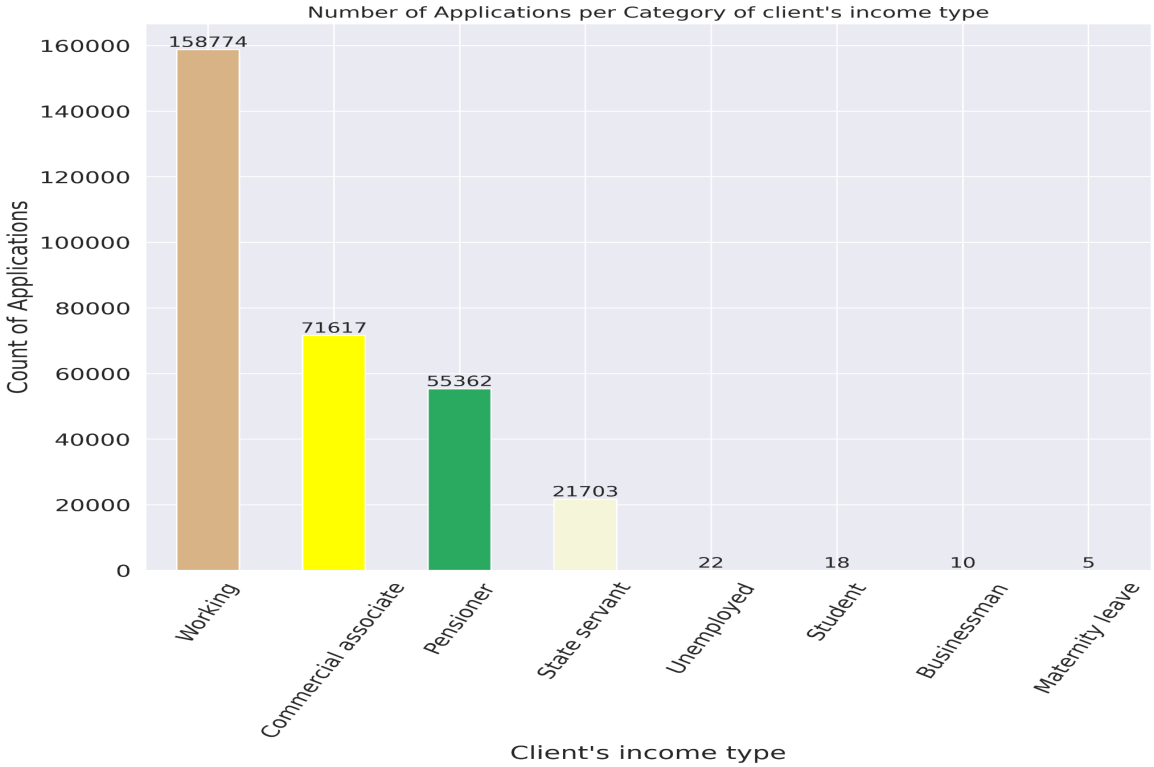
Key takeaway: These box plots of some continuous variables from the previous_application dataset represent outliers on the basis of quartile range. Out of 6, only 2 attributes have extreme outliers. These 2 are AMT_CREDIT, SELLERPLACE_AREA. Both of these attributes have no anomaly in outlier data and are possible in real world scenarios.

B. Understanding categorical variables:

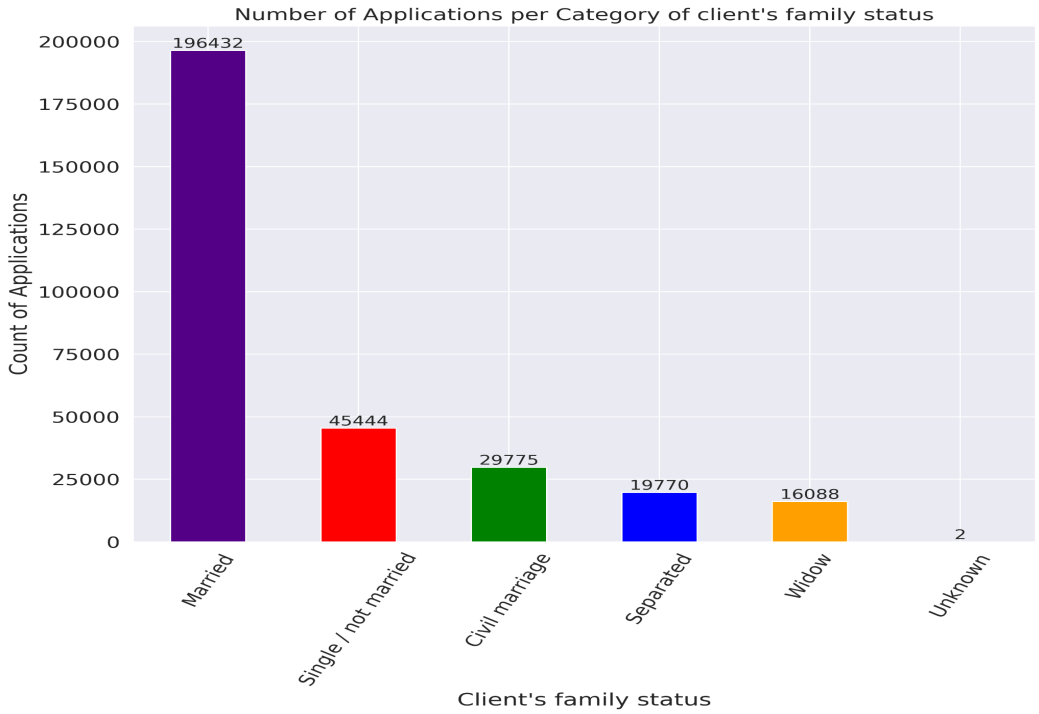
1. **Key Takeaway** - Around 81% applications received are from clients who were unaccompanied by family members.



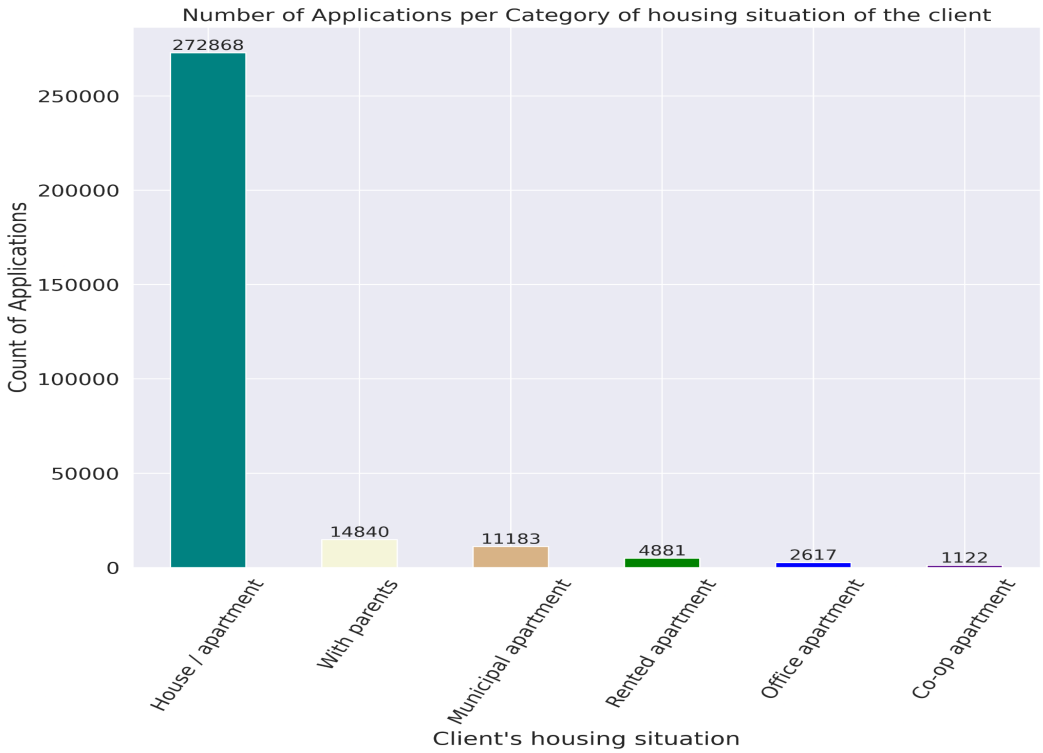
2. **Key Takeaway -** Around 51% applications received are from clients who were working.



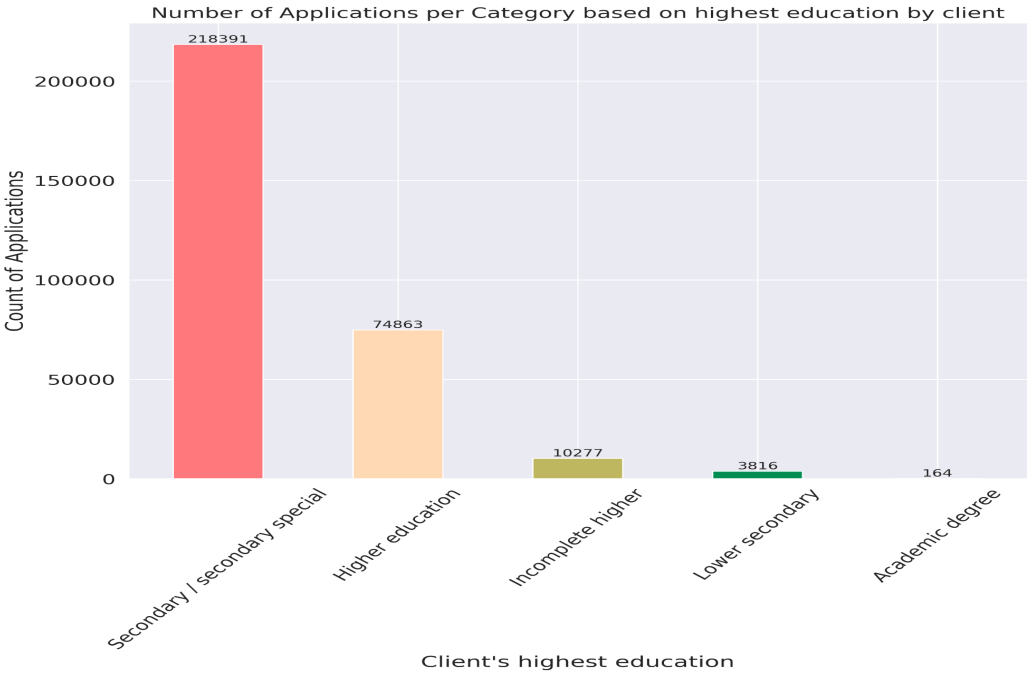
3. **Key Takeaway -** Around 73% applications received are from clients who were married including civil married



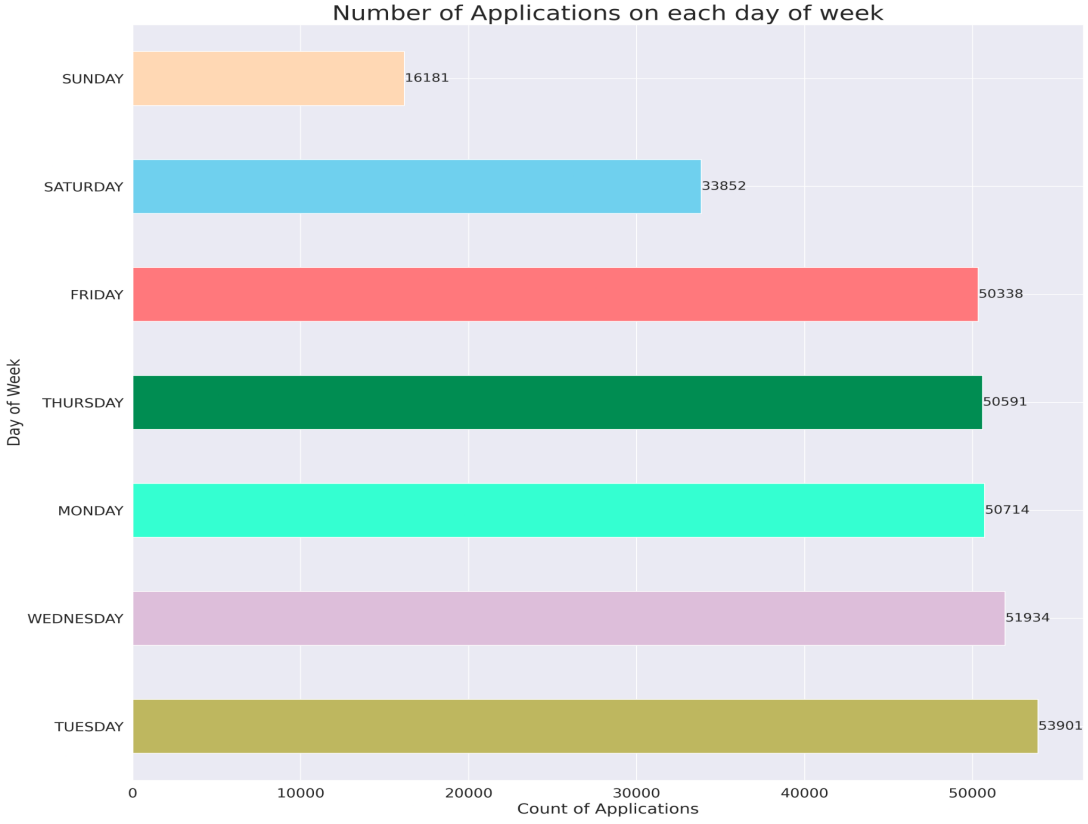
4. **Key Takeaway -** Around 88% applications received are from clients who were residing in a house/apartment.



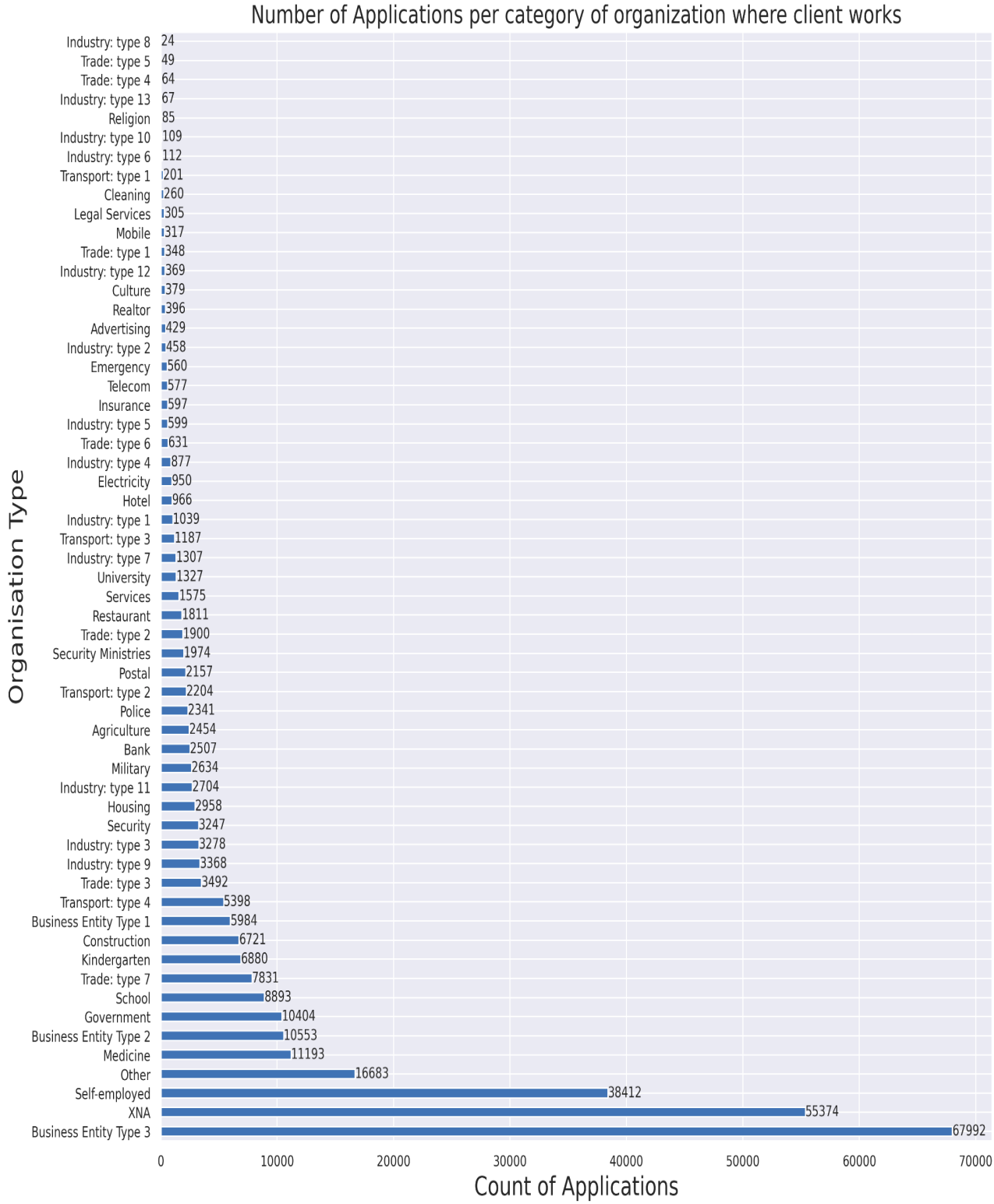
5. **Key Takeaway -** Around 71% applications received are from clients who were educated at secondary/secondary special level.



6. **Key Takeaway -** Around 17+-0.48% applications were received on weekdays from Monday to friday with tuesday being the day on which most applications were received. On weekends the applications received were less which is around 11% on saturdays and only 5% on sunday.

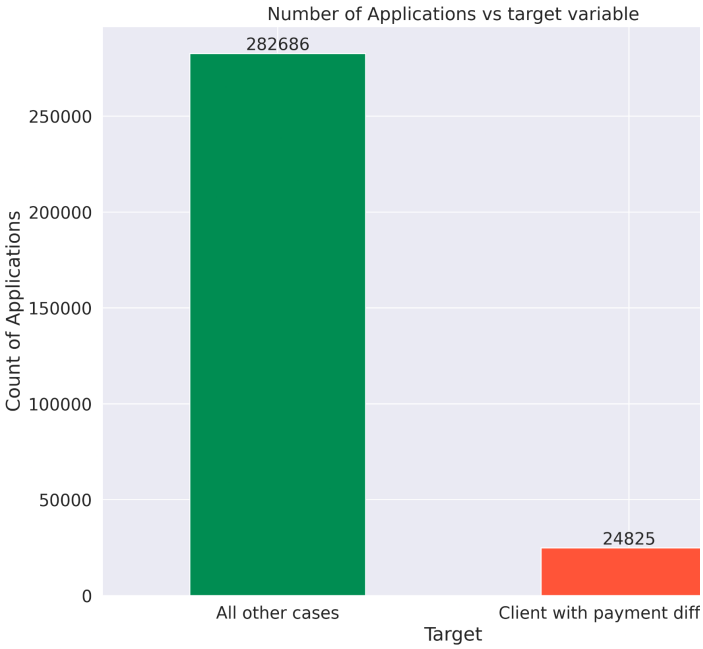


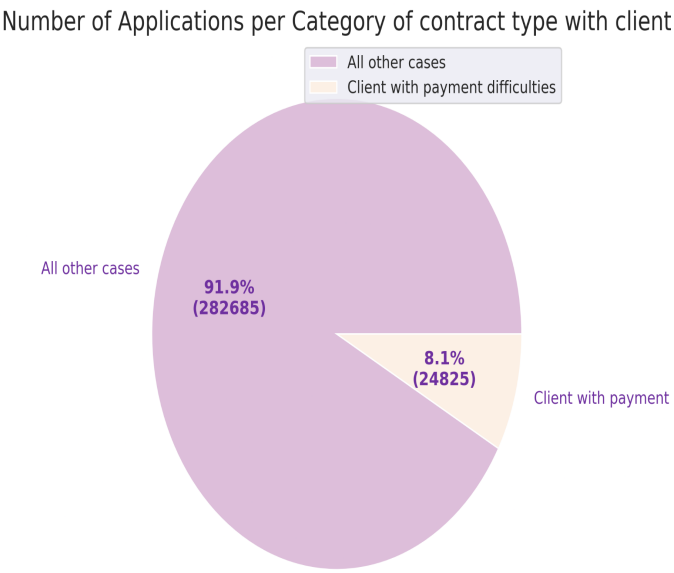
7. **Key Takeaway -** Majority Around 21% applications received are from clients who were working at business entity type 3 organization.



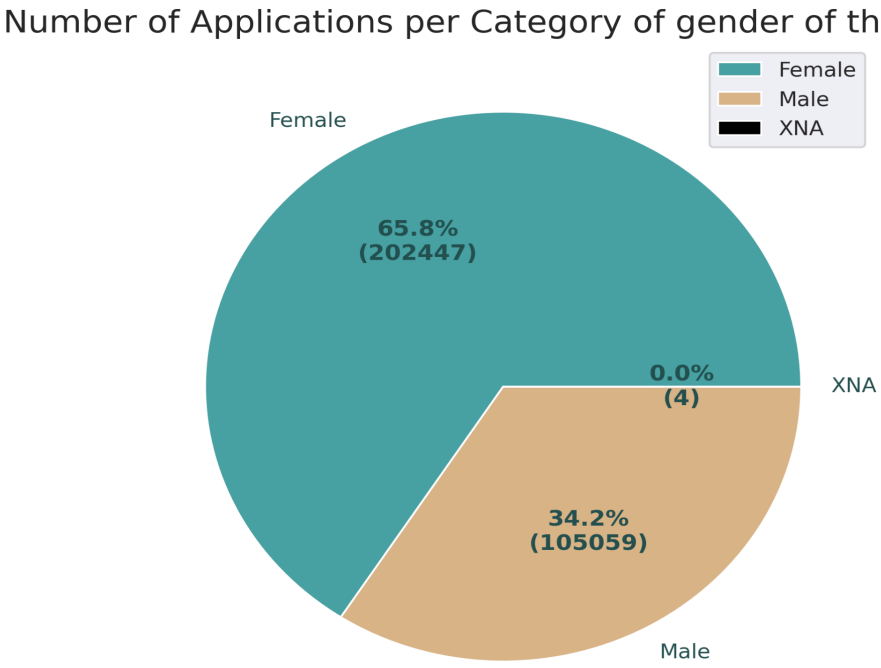
C. Data imbalance

1. **Key Takeaway:** Moderate degree of imbalance in data according to target variable

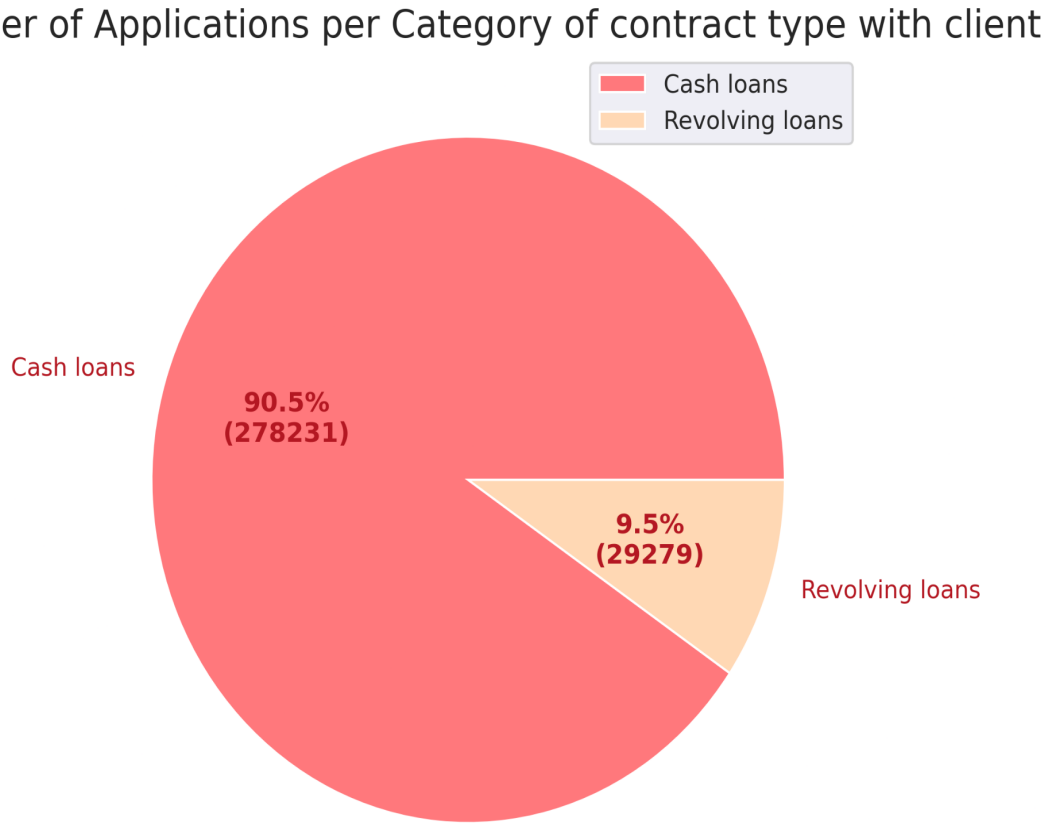




2. **Key Takeaway:** Mild degree of imbalance in data according to gender attribute



3. **Key Takeaway:** Moderate degree of imbalance in data according to contract type column

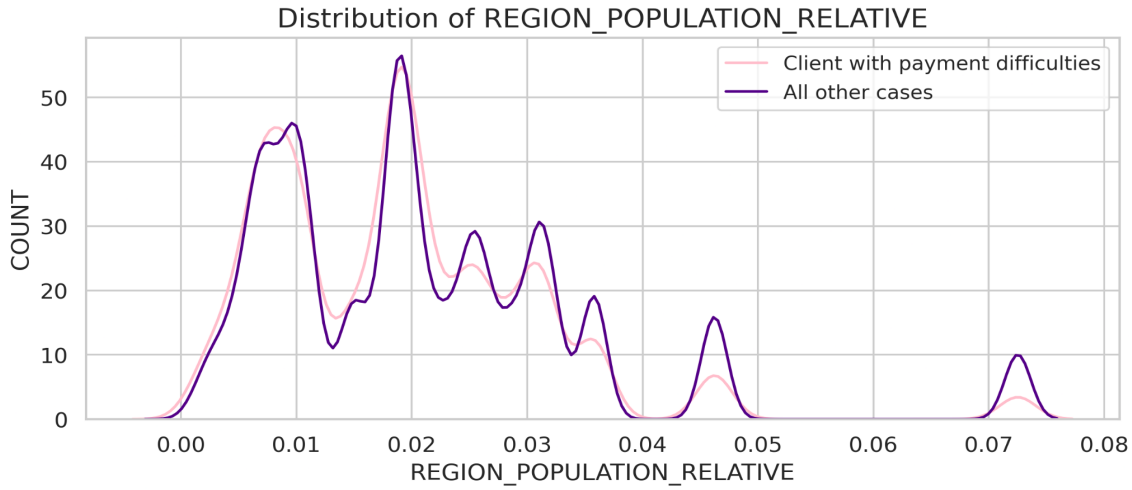


COLUMN NAME:	TARGET		
Attributes	0	1	
count	282686	24825	
Imbalance Ratio	91.93%	8.07%	
COLUMN NAME:	CODE_GENDER		
Attributes	F	M	XNA
count	202448	105059	4
Imbalance Ratio	65.83%	34.16%	0.00%
COLUMN NAME:	NAME_CONTRACT_TYPE		
Attributes	Cash loans	Revolving loans	
count	278232	29279	
Imbalance Ratio	90.48%	9.52%	

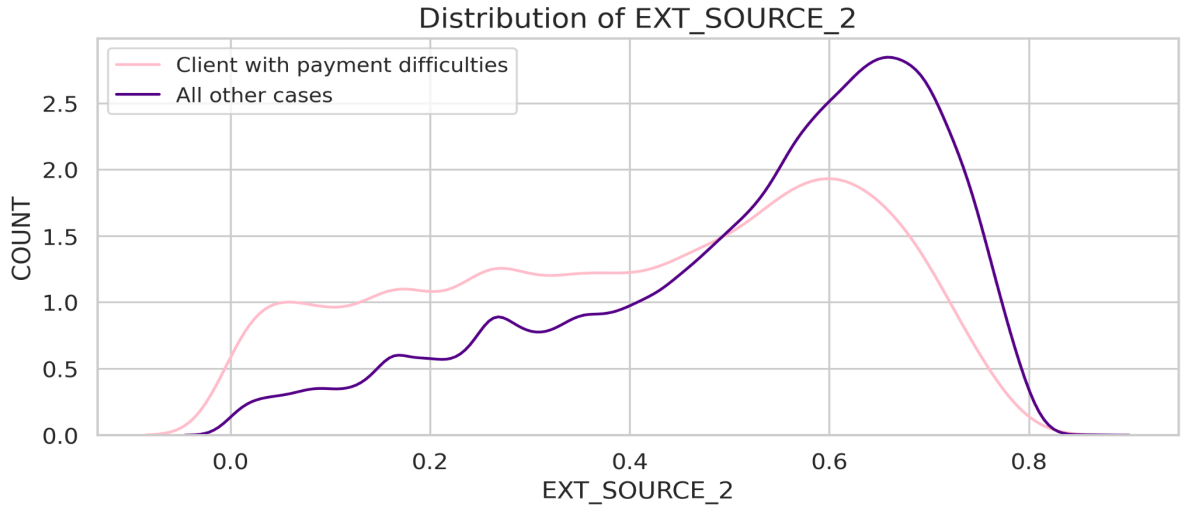
D. Univariate analysis:-

a. Numerical Variable analysis:

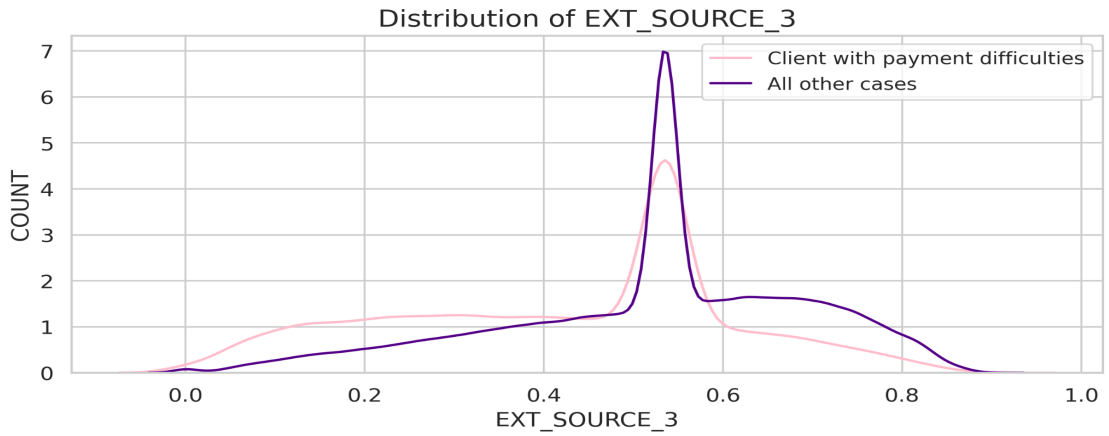
1. **Key Takeaway:** Most applications are from clients living in less populated region and also account for more clients with payment difficulties



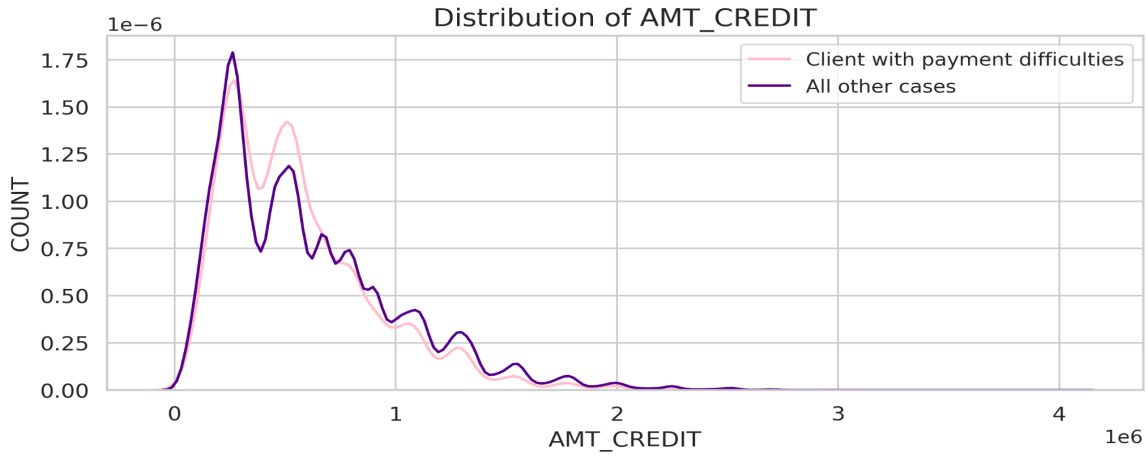
2. **Key Takeaway:** More number of applicants with payment difficulties have lower EXT_SOURCE_2 score and applicants who do not default have higher scores.



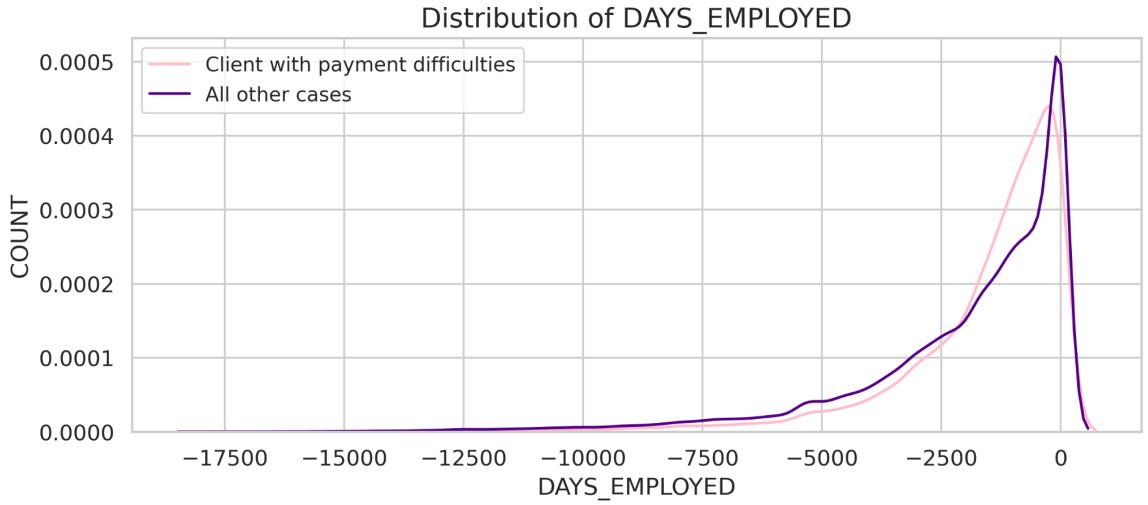
3. **Key Takeaway:** More number of applicants with payment difficulties have lower EXT_SOURCE_3 score and applicants who do not default have higher scores.



4. **Key Takeaway:** There is no significant difference in defaulter clients and non defaulter clients in case of credit amount.



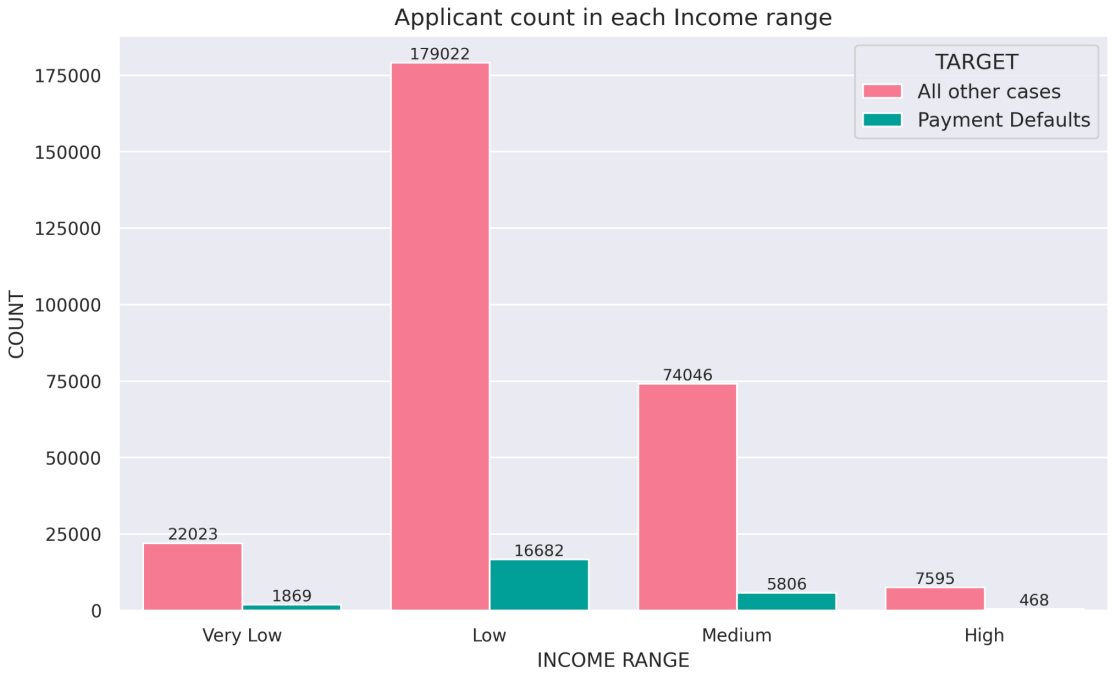
5. **Key Takeaway:** Applicants with low employment days tend to default more while applicants with more employment days do not have payment difficulties.



b. Categorical Variable analysis:

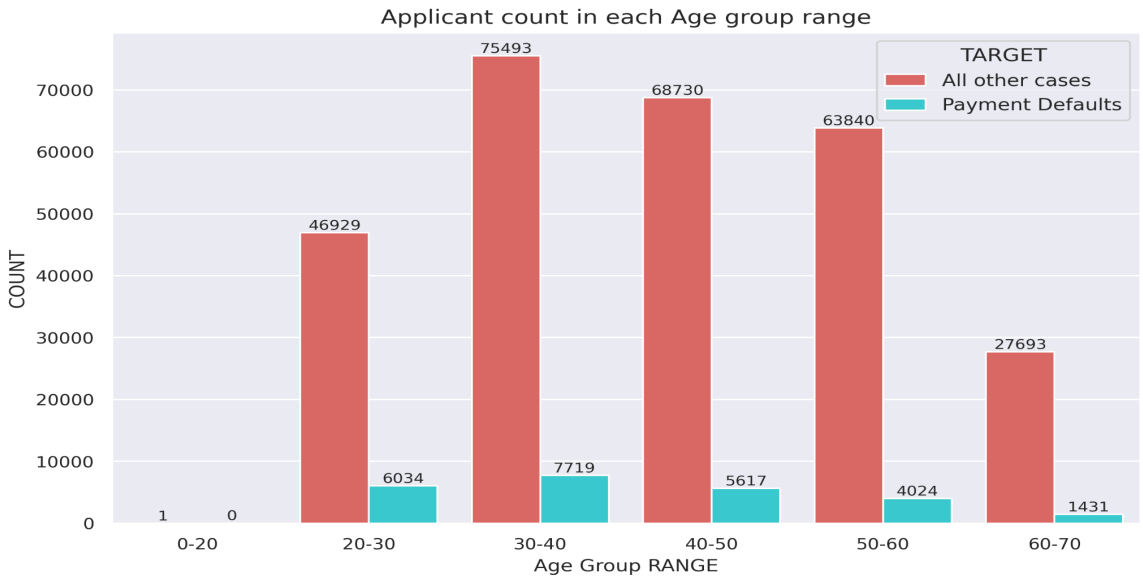
1. **Key Takeaway:** Applicants with high income range have less chances of payment difficulties.

AMT_INCOME_RANGE	All other Cases in different income range	Payment Defaults in different income range
Low	91.48%	8.52%
Medium	92.73%	7.27%
Very Low	92.18%	7.82%
High	94.20%	5.80%



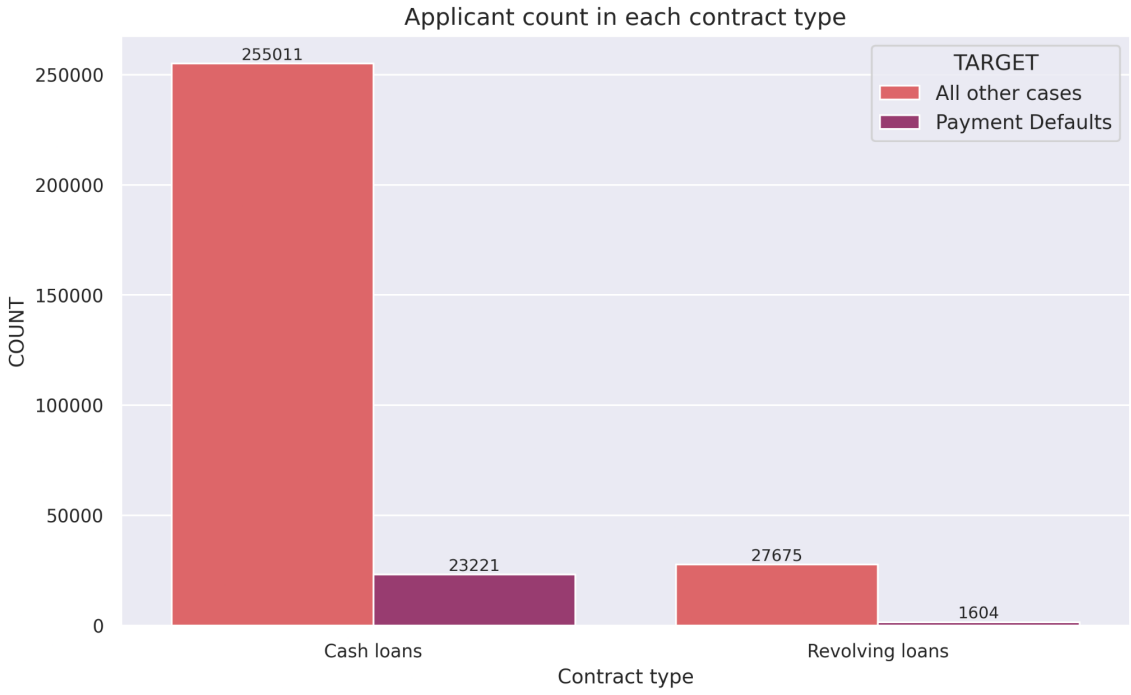
2. **Key Takeaway:** Applicants of age group 20-30 and 30-40 yrs have high chances of payment difficulties.

AGE_RANGE	Payment Defaults in different age range	All other Cases in different age range
0-20	0.00%	100.00%
20-30	11.39%	88.61%
30-40	9.28%	90.72%
40-50	7.56%	92.44%
50-60	5.93%	94.07%
60-70	4.91%	95.09%



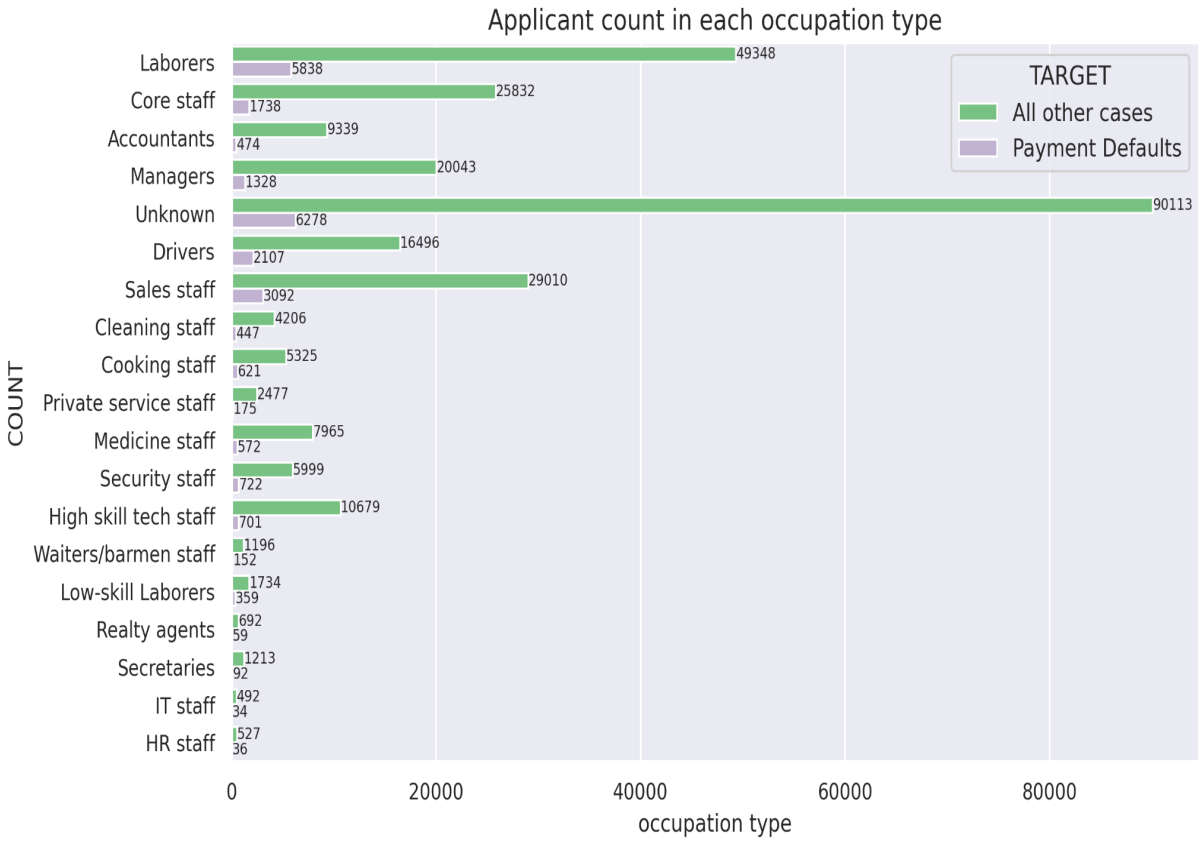
3. **Key Takeaway:** Application of cash loans have high chances of payment difficulties than of revolving loans.

NAME_CONTRACT_TYPE	Payment Defaults in each contract type	All other Cases in each contract type
Cash Loans	8.35%	91.65%
Revolving Loans	5.48%	94.52%



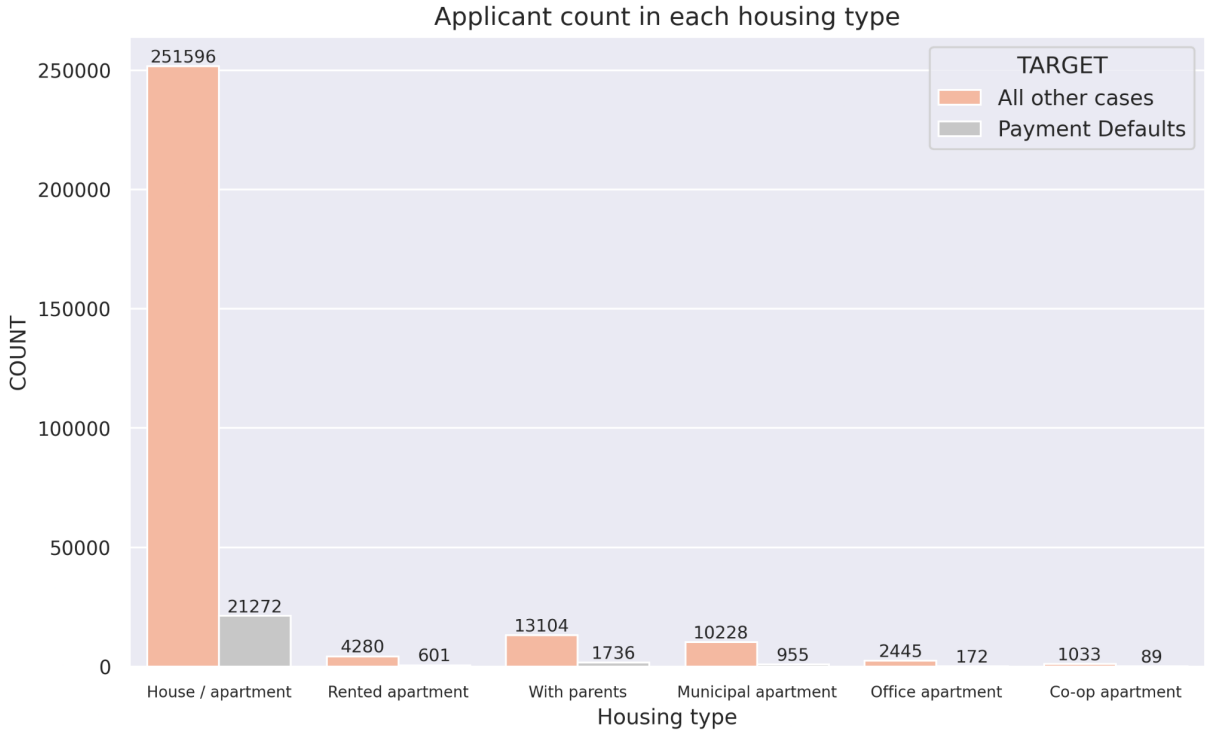
4. **Key Takeaway:** Applicants who are low-skilled laborers have highest chances of payment difficulties and applicants who are high skill tech staff and accountants have less cases of payment defaults.

OCCUPATION_TYPE	Payment Defaults in each occupation type	All other Cases in each occupation type
Accountants	4.83%	95.17%
Cleaning staff	9.61%	90.39%
Cooking staff	10.44%	89.56%
Core staff	6.30%	93.70%
Drivers	11.33%	88.67%
HR staff	6.39%	93.61%
High skill tech staff	6.16%	93.84%
IT staff	6.46%	93.54%
Laborers	10.58%	89.42%
Low-skill Laborers	17.15%	82.85%
Managers	6.21%	93.79%
Medicine staff	6.70%	93.30%
Private service staff	6.60%	93.40%
Realty agents	7.86%	92.14%
Sales staff	9.63%	90.37%
Secretaries	7.05%	92.95%
Security staff	10.74%	89.26%
Unknown	6.51%	93.49%
Waiters/barmen staff	11.28%	88.72%



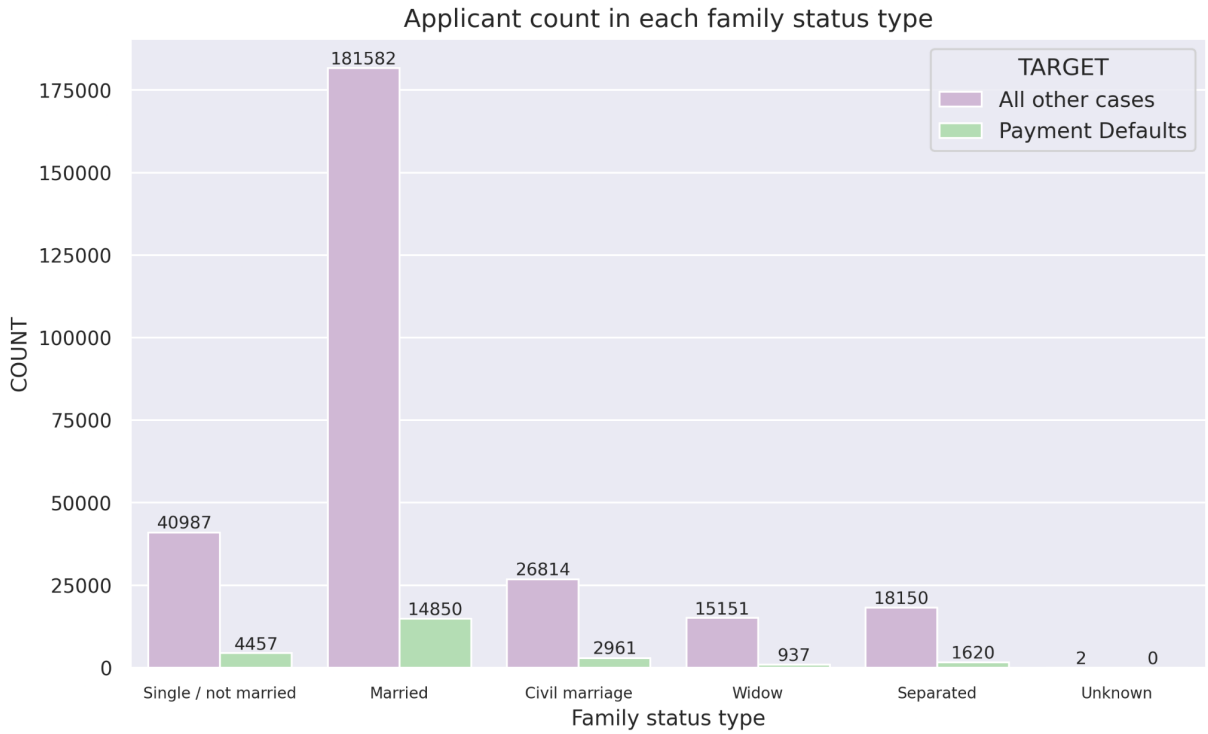
5. **Key Takeaway:** Applicants living with parents or in rented apartments have more chances of payment difficulties than applicants living in other housing types.

Housing type	Payment Defaults in each housing type	All other Cases in each housing type
House / apartment	7.80%	92.20%
With parents	11.70%	88.30%
Municipal apartment	8.54%	91.46%
Rented apartment	12.31%	87.69%
Office apartment	6.57%	93.43%
Co-op apartment	7.93%	92.07%



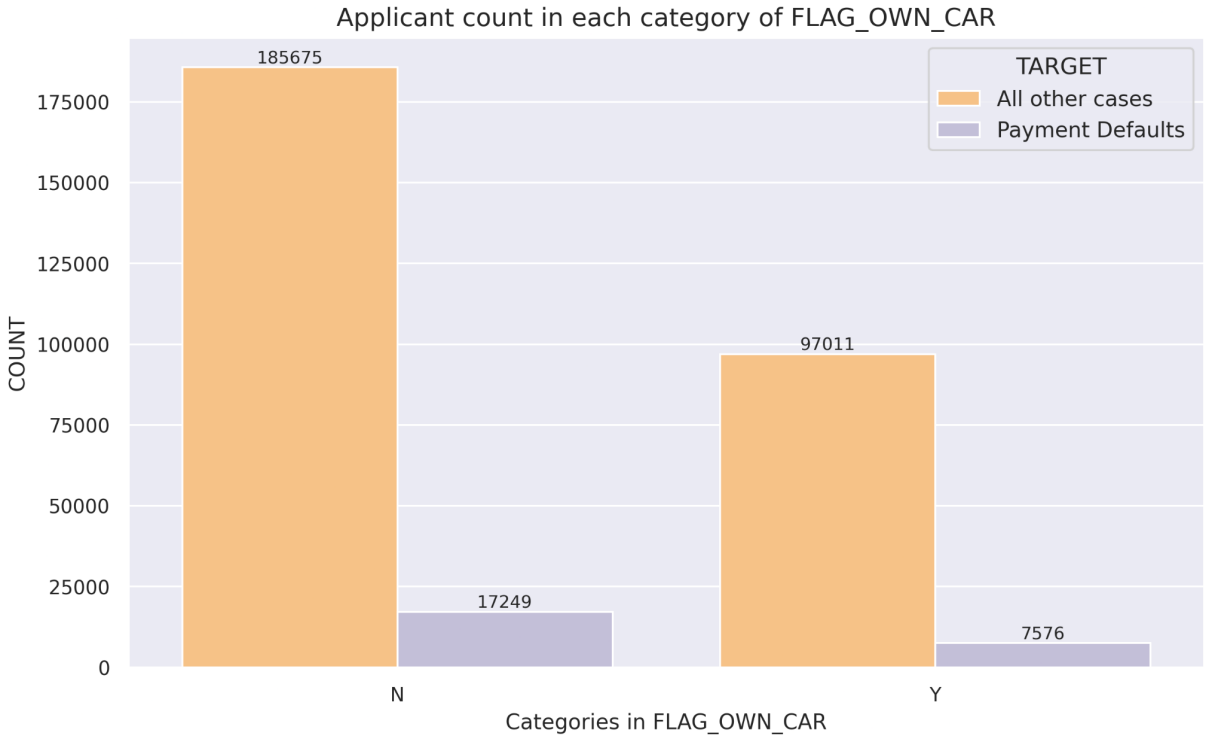
6. **Key Takeaway:** Applicants who are widow have less chances of payment defaults where as people who are civil married have the highest.

NAME_FAMILY_STATUS	Payment Defaults in different age range	All other Cases in different age range
Civil marriage	9.94%	90.06%
Married	7.56%	92.44%
Separated	8.19%	91.81%
Single / not married	9.81%	90.19%
Unknown	0.00%	100.00%
Widow	5.82%	94.18%



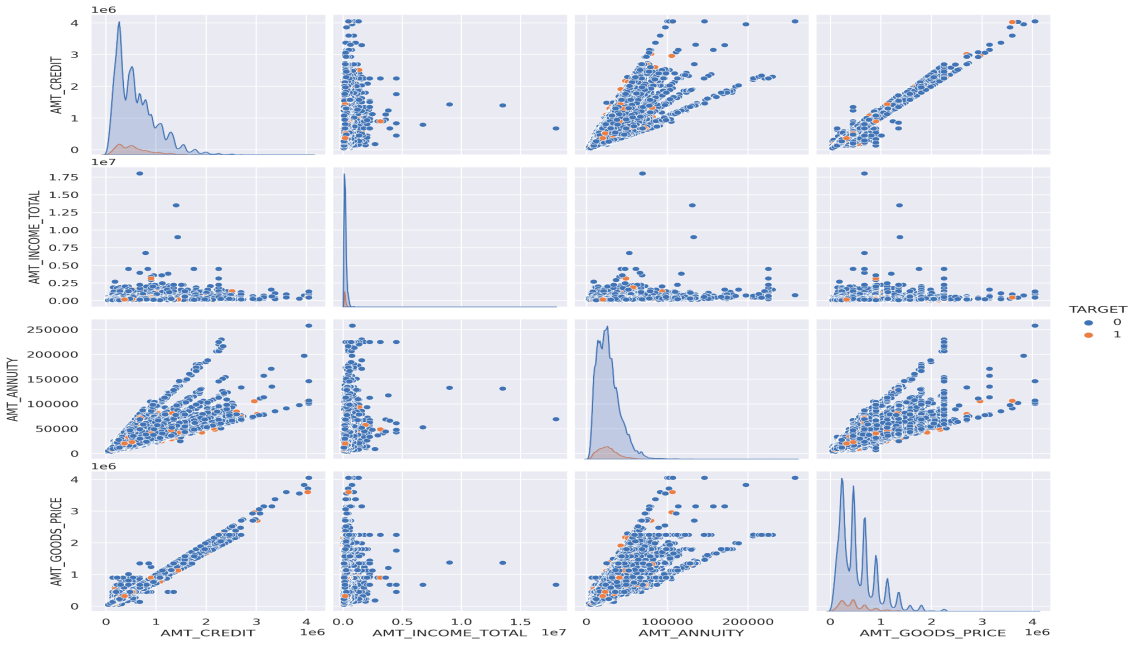
7. **Key Takeaway:** People who do not own a car have faced payment difficulties.

FLAG_OWN_CAR	Payment Defaults	All other Cases
N	8.50%	91.50%
Y	7.24%	92.76%

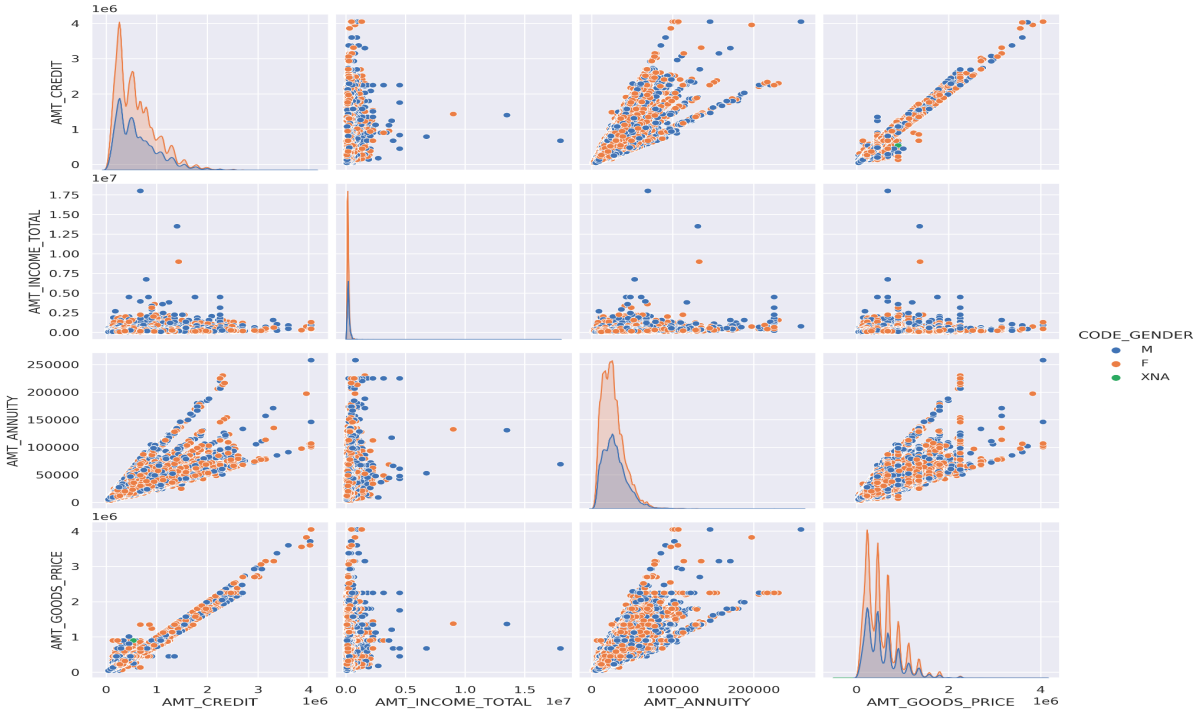


E. Bivariate analysis:-

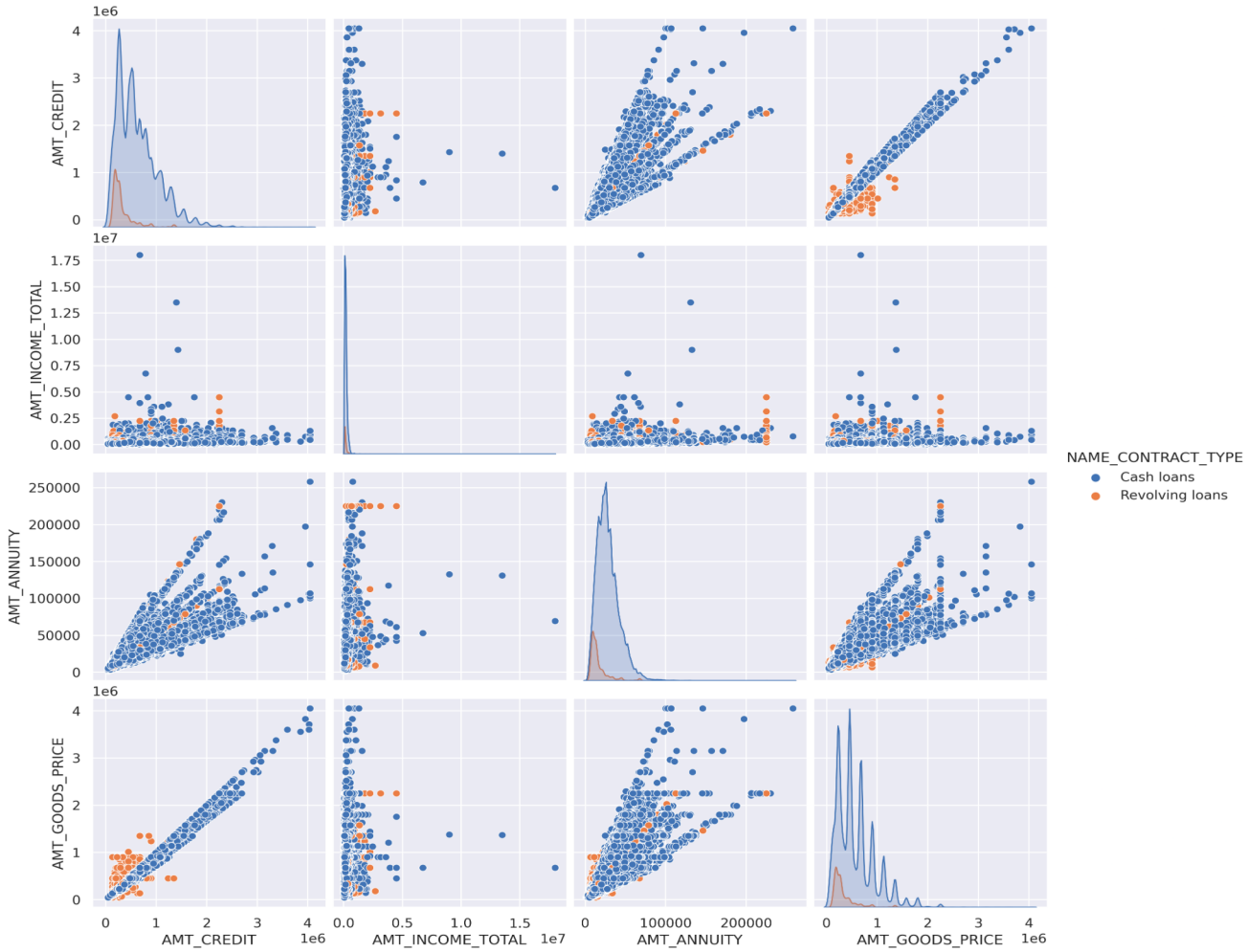
1. Bivariate analysis of 4 continuous variables AMT_CREDIT, AMT_INCOME_TOTAL, AMT_ANNUITY, AMT_GOODS_PRICE categorized on the basis of target attribute.



Bivariate analysis of 4 continuous variables AMT_CREDIT, AMT_INCOME_TOTAL, AMT_ANNUITY, AMT_GOODS_PRICE categorized on the basis of code_gender attribute.

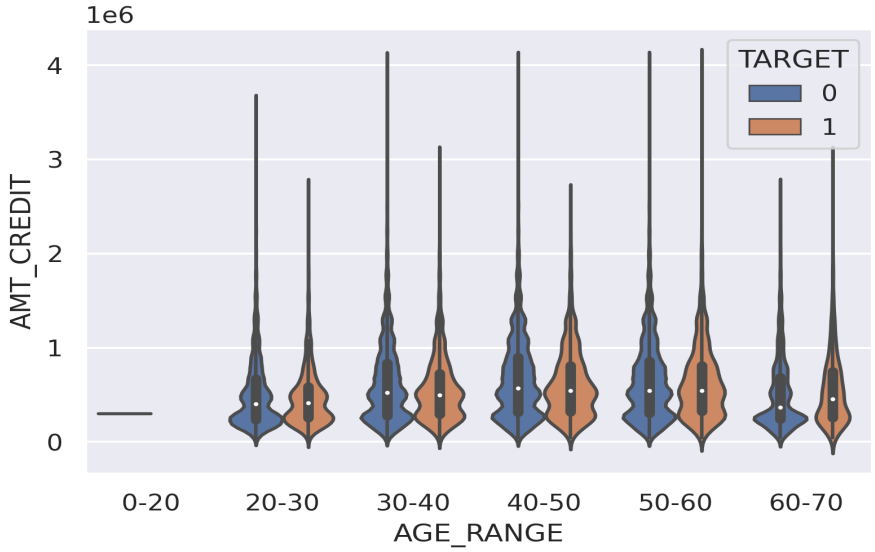


Bivariate analysis of 4 continuous variables AMT_CREDIT, AMT_INCOME_TOTAL, AMT_ANNUITY, AMT_GOODS_PRICE categorized on the basis of name_contract_type attribute.

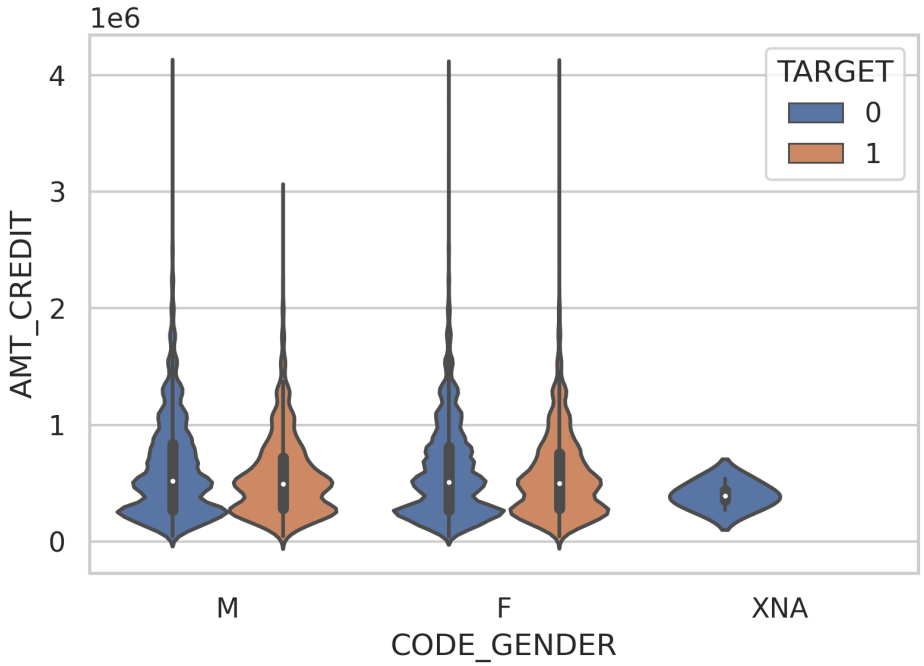


Key Takeaway: There is correlation between AMT_CREDIT, AMT_GOODS_PRICE and AMT_ANNUITY. There is no direct relation between AMT_INCOME_TOTAL and the other three attributes.

2. Violin Plot illustrating distribution of AMT_CREDIT vs AGE_RANGE and AMT_CREDIT vs CODE_GENDER with respect to TARGET.

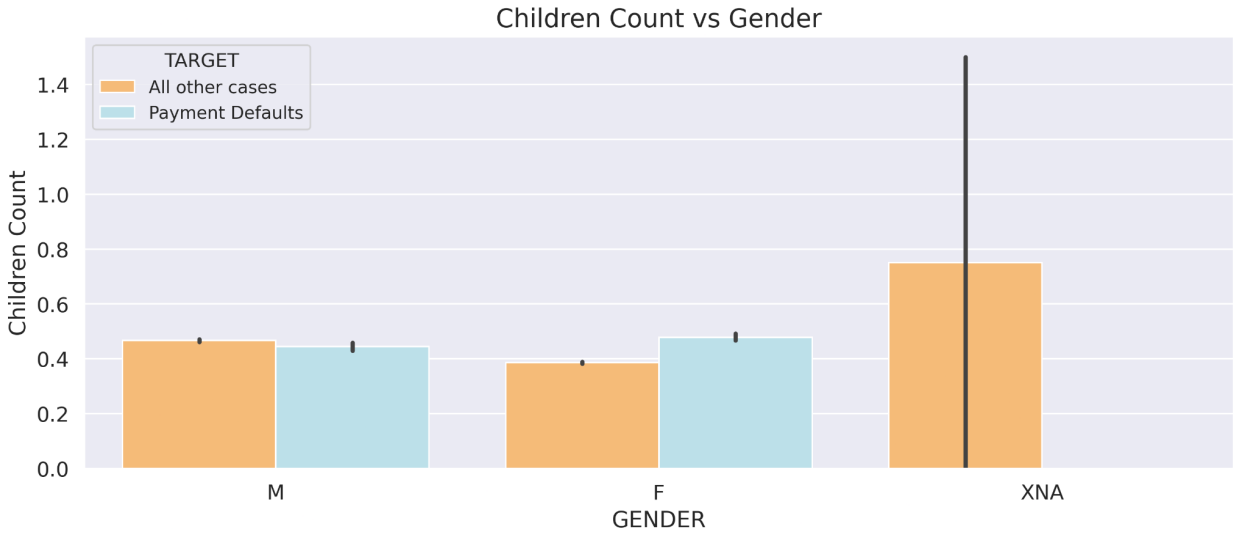


Key Takeaway: From the above plot, it is clearly visible that in each age range from 20-70, applicants applied for credit amounts of 0-10 lakh in most cases.



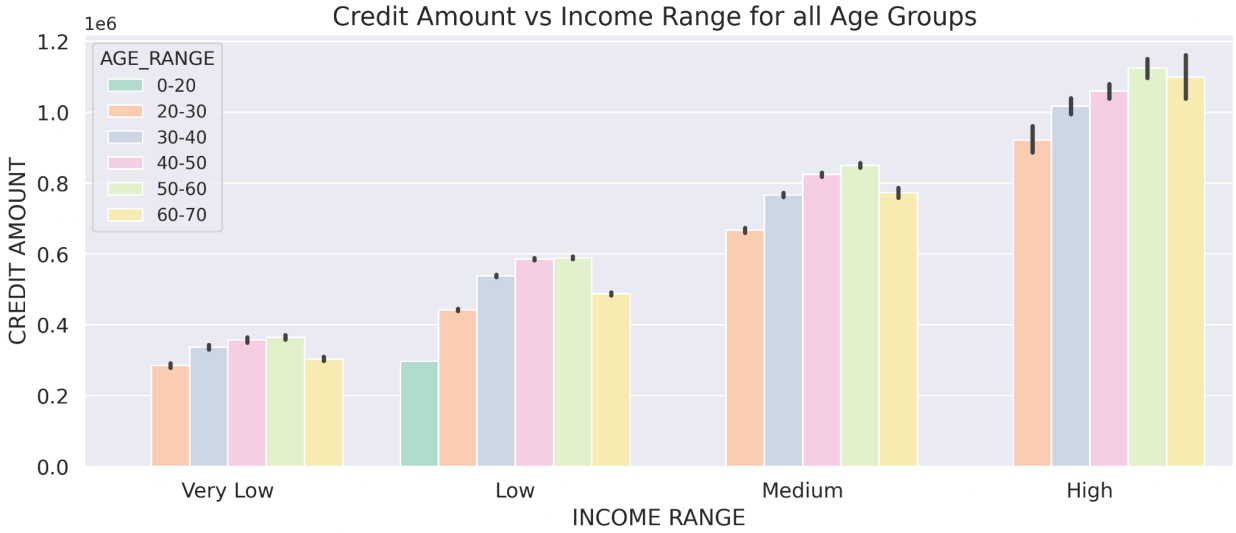
Key Takeaway: From the above plot, it is clearly visible that both male and female have applied for a maximum of around 40 lakhs of credit. The males who defaulted on payment have applied for a maximum of around 30 lakhs of credit amount which is less than that of females.

3. Barplot below illustrates the relation between children count and code gender attribute.



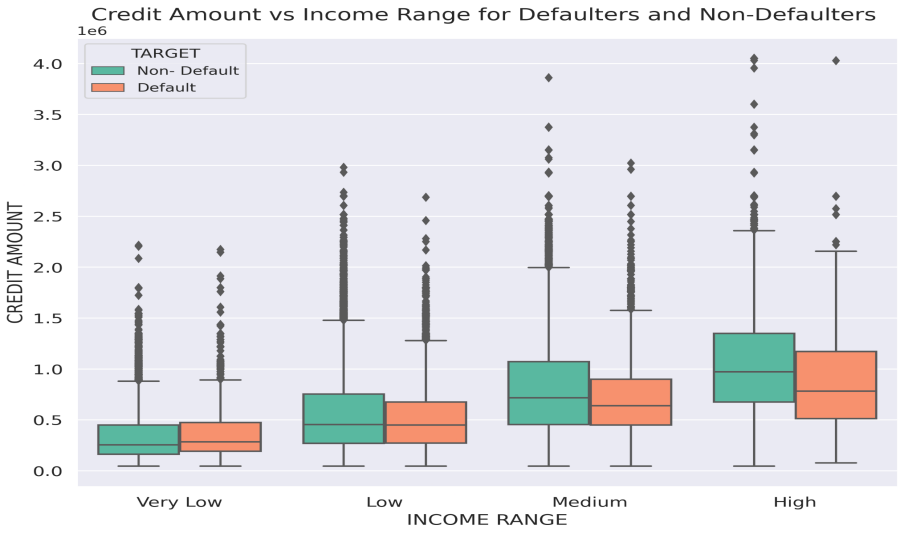
Key Takeaway: From the above plot, it is clearly visible that females with more children have chances of payment difficulties. In contrast to this, male with fewer children have chances of payment difficulties.

4. Barplot below illustrates the multivariate relation between credit amount ,income range and age range attribute.



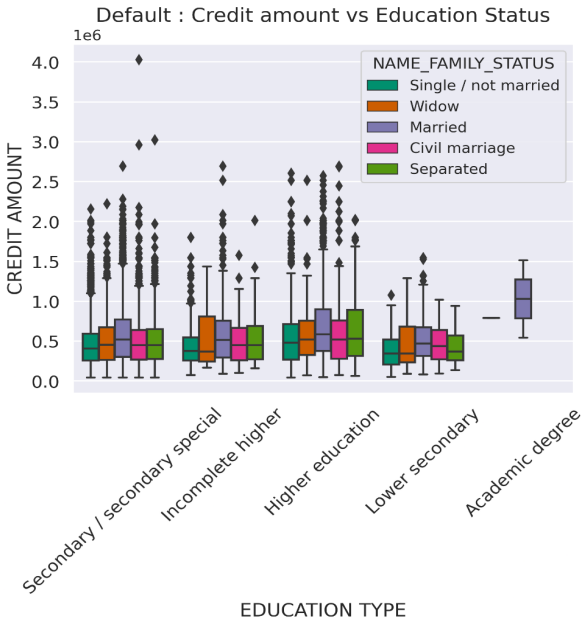
Key Takeaway: From the above plot, it is clearly visible that credit amount is directly proportional with income range and the pattern of age range with respect to credit amount in each income group remains the same where the credit amount increases from age of 20-60 and drops in 60-70 range.

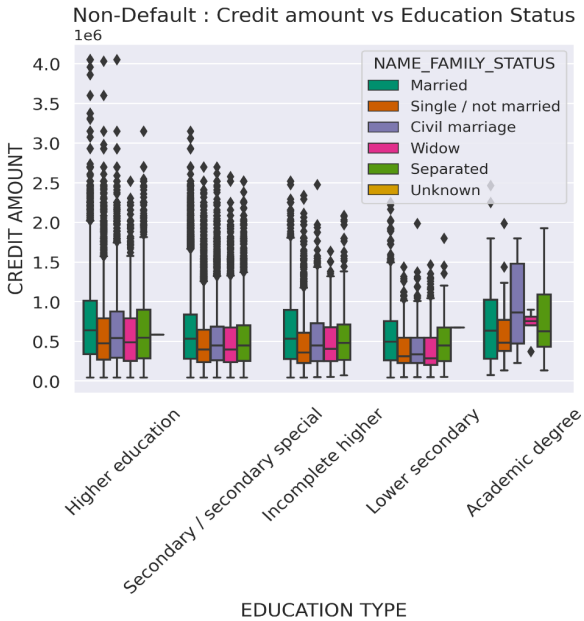
5. Boxplot below illustrates the multivariate relation between credit amount ,income range and target attribute.



Key Takeaway: From the above plot, it is clearly visible that applicants in the very low income range with an annual salary less than 75000 have difficulty in payments when credit amount is high. Rest applicants in other 3 income range groups with income shows a similar pattern. All of the applicants in these income ranges have payment difficulties in case of lesser credit amount.

6. Boxplot below illustrates the multivariate relation between credit amount ,income range and target attribute.

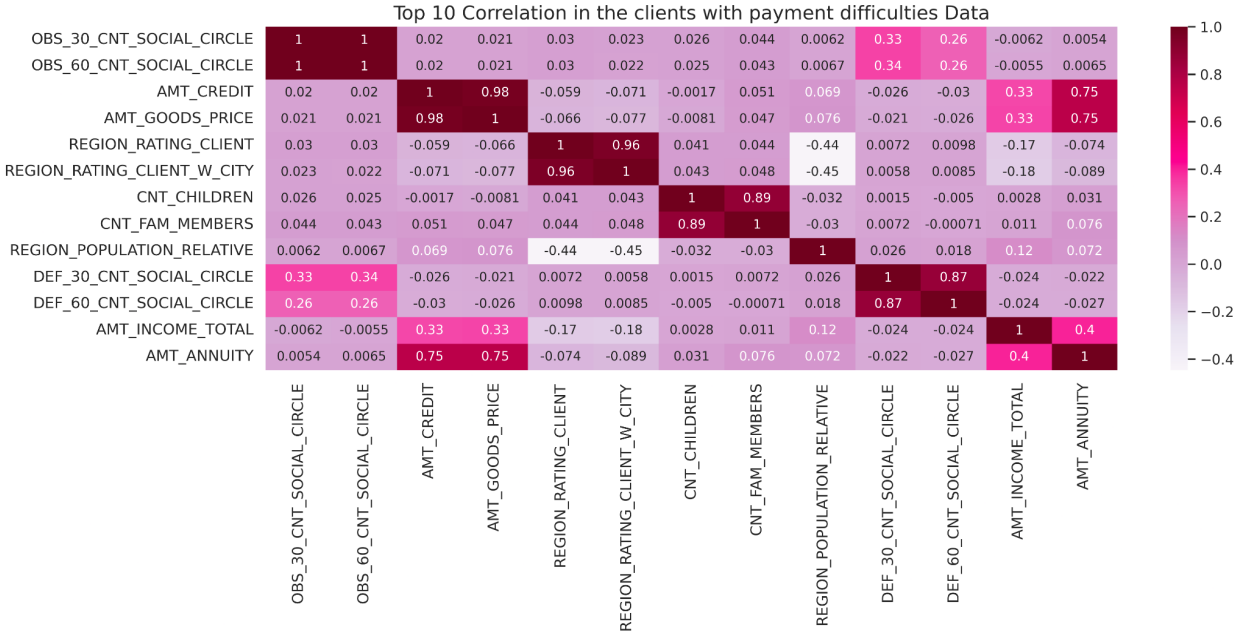




Key Takeaway: From the above plot, it is clearly visible that applicants with high education have applied for large credit amounts and also defaulted more on payments. Out of all applicants who have academic degrees only people who are married defaulted on payments.

F. Top 10 correlations :-

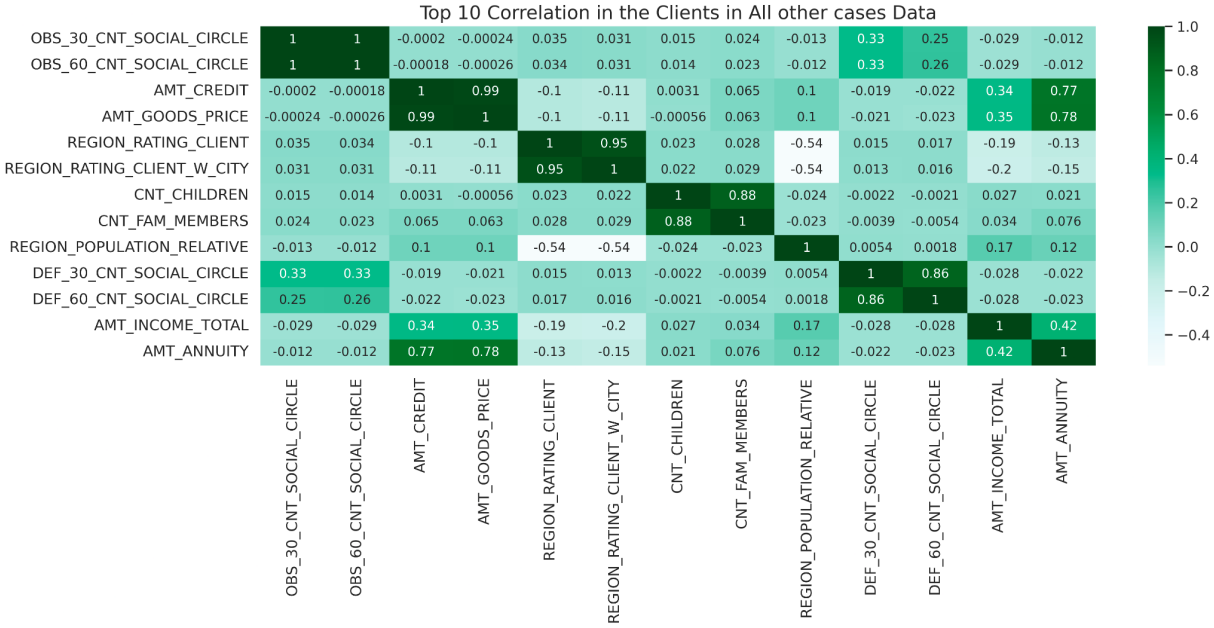
1. Heatmap below illustrates the top 10 correlation in application data of clients who have payment difficulties.



Key Takeaway: From the above plot, top 10 correlations in application data of clients who have payment difficulties are clearly visible with the correlation score and the results are presented in table below.

Attribute-1	Attribute-2	Correlation
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	1.00
AMT_CREDIT	AMT_GOODS_PRICE	0.98
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.96
CNT_CHILDREN	CNT_FAM_MEMBERS	0.89
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.87
AMT_ANNUITY	AMT_GOODS_PRICE	0.75
AMT_CREDIT	AMT_ANNUITY	0.75
AMT_INCOME_TOTAL	AMT_ANNUITY	0.40
DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.34
AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.33

2. Heatmap below illustrates the top 10 correlation in application data of clients in all other cases.



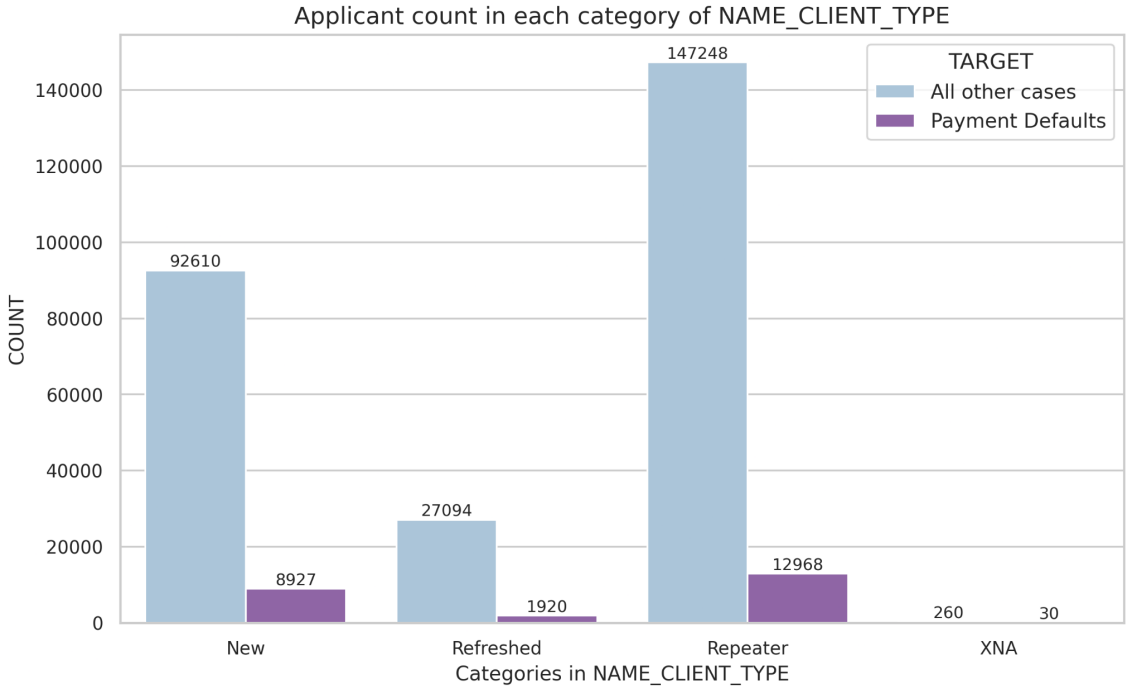
Key Takeaway: From the above plot, top 10 correlations in application data of clients in all other cases are clearly visible with the correlation score and the results are presented in the table below.

Attribute-1	Attribute-2	Correlation
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	1.00
AMT_CREDIT	AMT_GOODS_PRICE	0.99
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.95
CNT_CHILDREN	CNT_FAM_MEMBERS	0.88
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.86
AMT_ANNUITY	AMT_GOODS_PRICE	0.78
AMT_CREDIT	AMT_ANNUITY	0.77
AMT_INCOME_TOTAL	AMT_ANNUITY	0.42
AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.35
AMT_INCOME_TOTAL	AMT_CREDIT	0.34

F. Analysis of Merged application_data and previous_application dataset :-

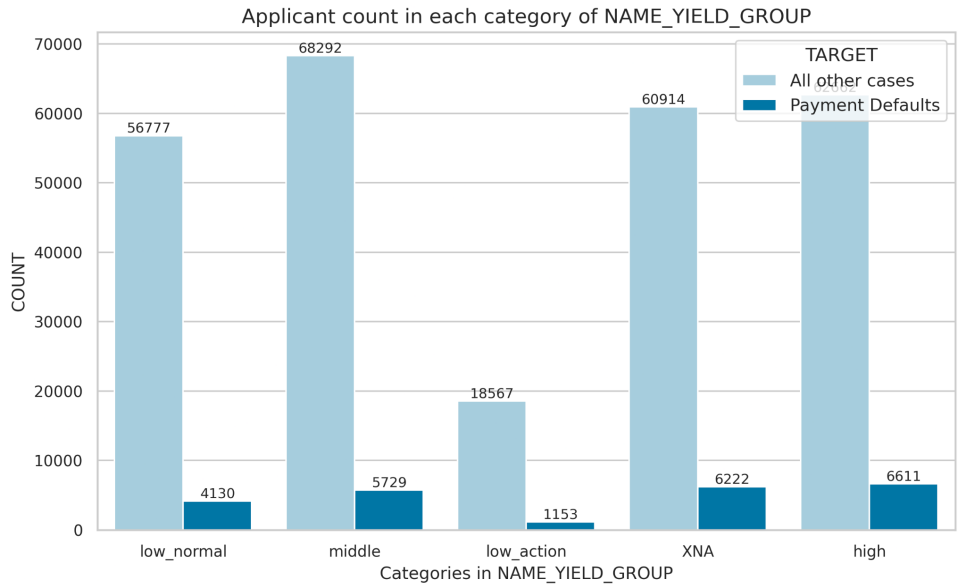
1. **Key Takeaway:** Applicants who are refreshed have less cases of payment defaults where as who are new have more cases of payment defaults.

NAME_CLIENT_TYPE	Payment defaulters by each client type	All other Cases by each client type
Repeater	8.09%	91.91%
New	8.79%	91.21%
Refreshed	6.62%	93.38%
XNA	10.34%	89.66%



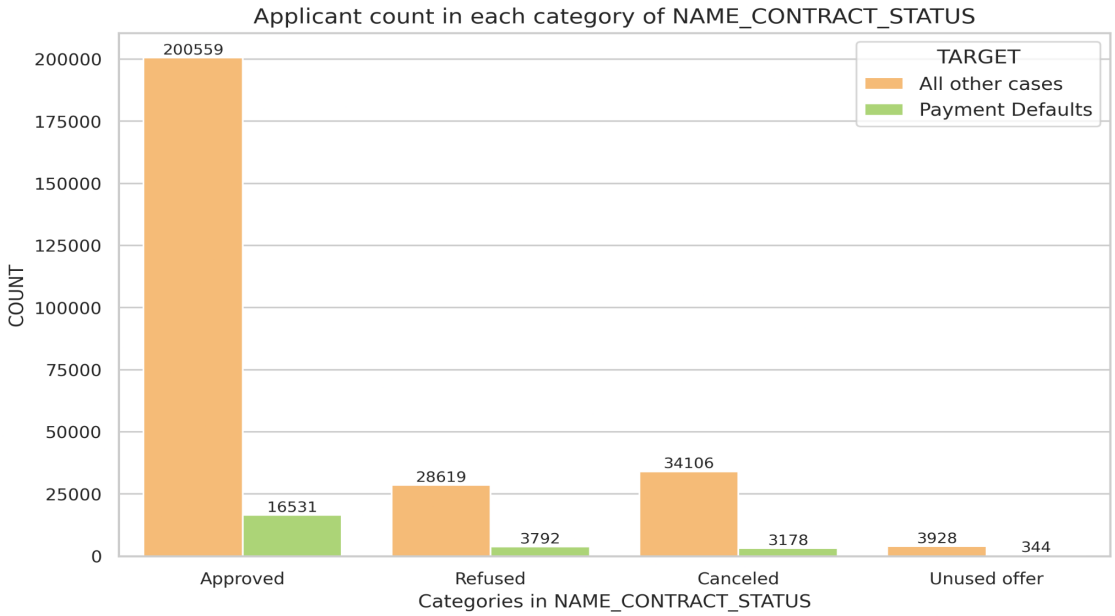
2. **Key Takeaway:** Applicants whose previous application interest rate are grouped into high yield group have more cases of payment difficulties where as same in low_action yield group have less chance of payment difficulties. Here, XNA represents clients with no information about previous application interest rates and also have higher cases of payment difficulties.

NAME_YIELD_GROUP	Payment defaulters by each yield group	All other Cases by each yield group
XNA	9.27%	90.73%
high	9.54%	90.46%
low_action	5.85%	94.15%
low_normal	6.78%	93.22%
middle	7.74%	92.26%



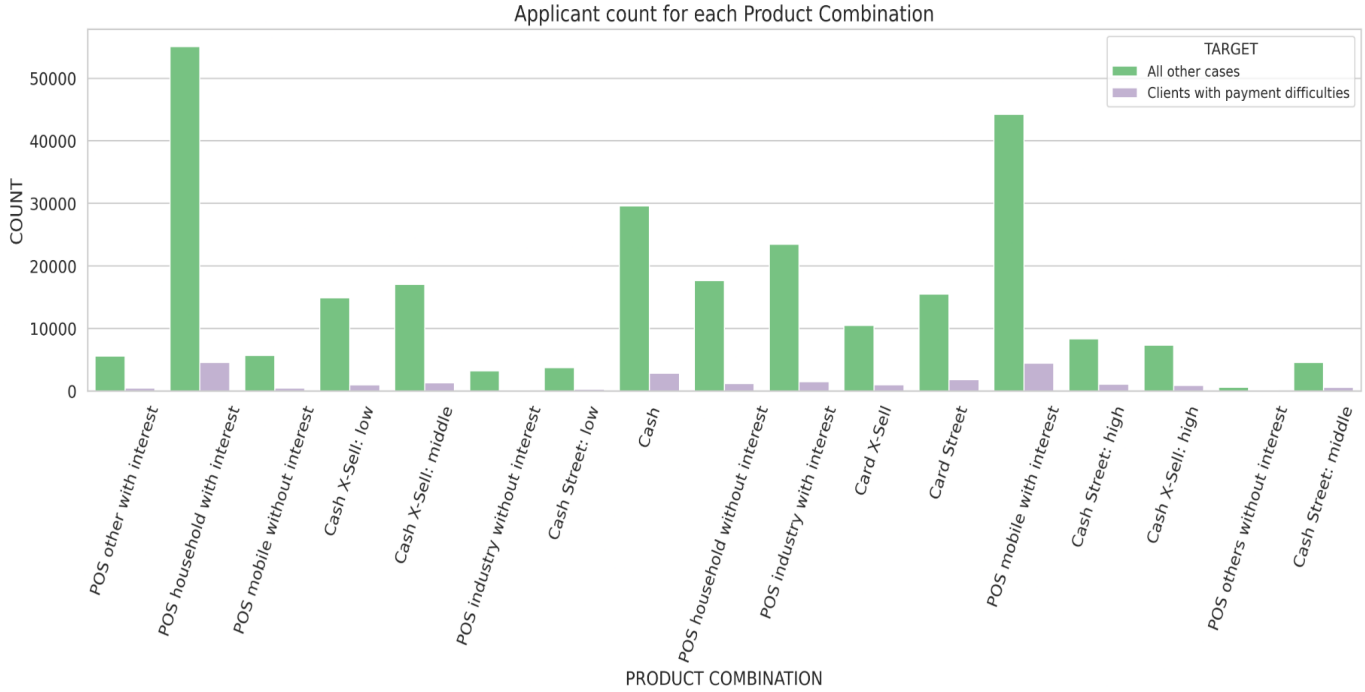
3. **Key Takeaway:** Applicants whose previous applications are approved have less chance of payment defaults and same of whose are refused have more cases of payment defaults.

NAME_CONTRACT_STATUS	Clients with payment difficulties by each contract type	All other Cases by each contract type
Approved	7.61%	92.39%
Canceled	8.52%	91.48%
Refused	11.70%	88.30%
Unused offer	8.05%	91.95%

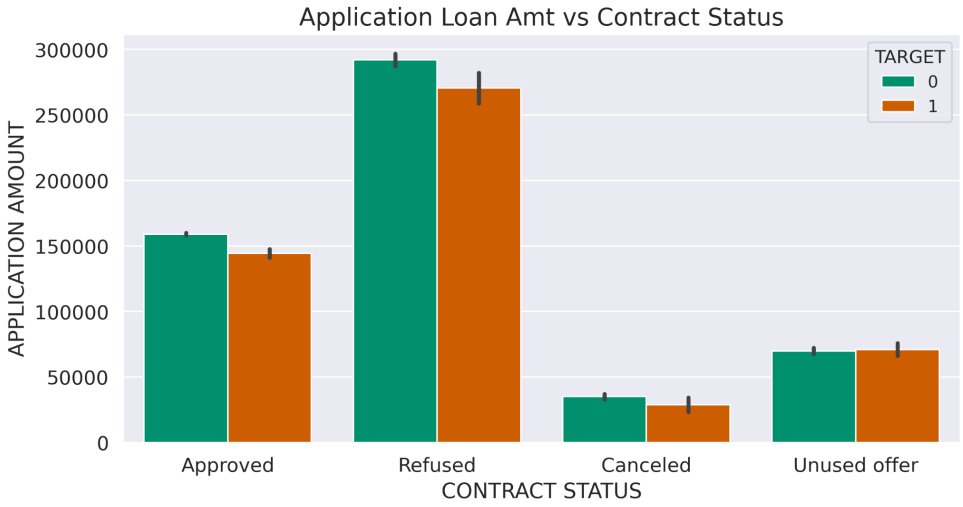


4. **Key Takeaway:** Applicants who take credit at POS for industry without interest have least cases of payment defaults. Applicants with cash product combination overall have more payment defaults. Also, with the presence of interest on credit, the default rate cases increases.

PRODUCT_COMBINATION	Clients with payment difficulties by each product combination	All other Cases by each product combination
Card Street	10.73%	89.27%
Card X-Sell	8.81%	91.19%
Cash	8.78%	91.22%
Cash Street: high	11.59%	88.41%
Cash Street: low	8.38%	91.62%
Cash Street: middle	11.23%	88.77%
Cash X-Sell: high	11.31%	88.69%
Cash X-Sell: low	6.29%	93.71%
Cash X-Sell: middle	7.40%	92.60%
POS household with interest	7.68%	92.32%
POS household without interest	6.20%	93.80%
POS industry with interest	5.90%	94.10%
POS industry without interest	3.67%	96.33%
POS mobile with interest	9.13%	90.87%
POS mobile without interest	7.94%	92.06%
POS other with interest	7.64%	92.36%
POS others without interest	8.77%	91.23%

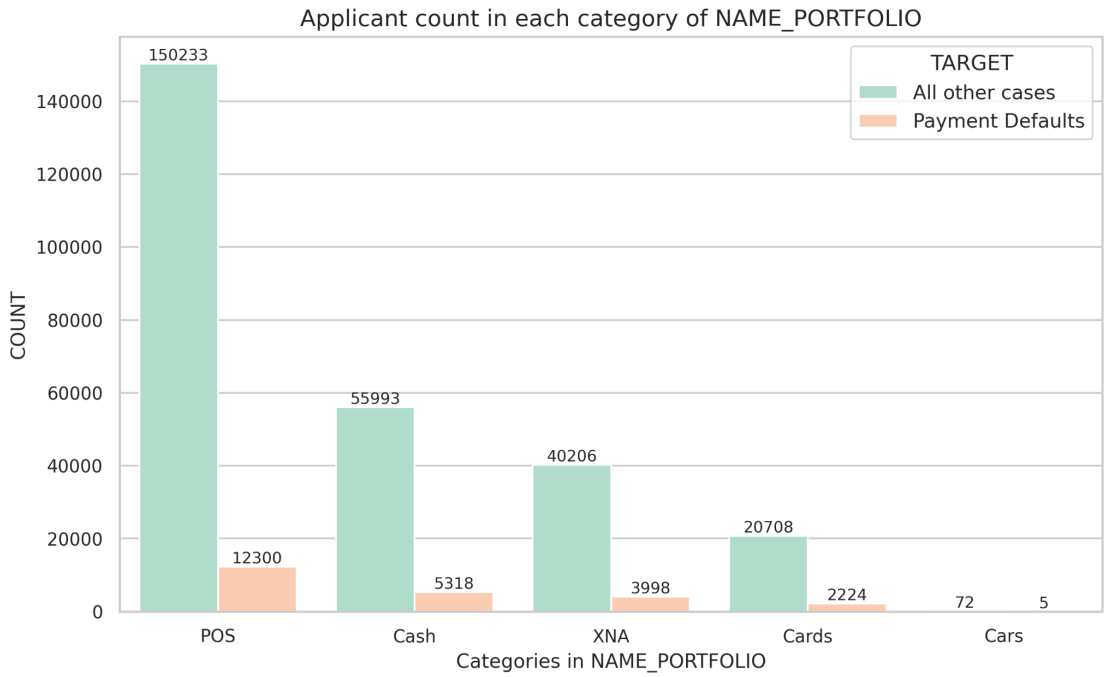


5. **Key Takeaway:** Applicants who applied for more than 1.75 lakhs are refused, even some clients have no payment difficulties.



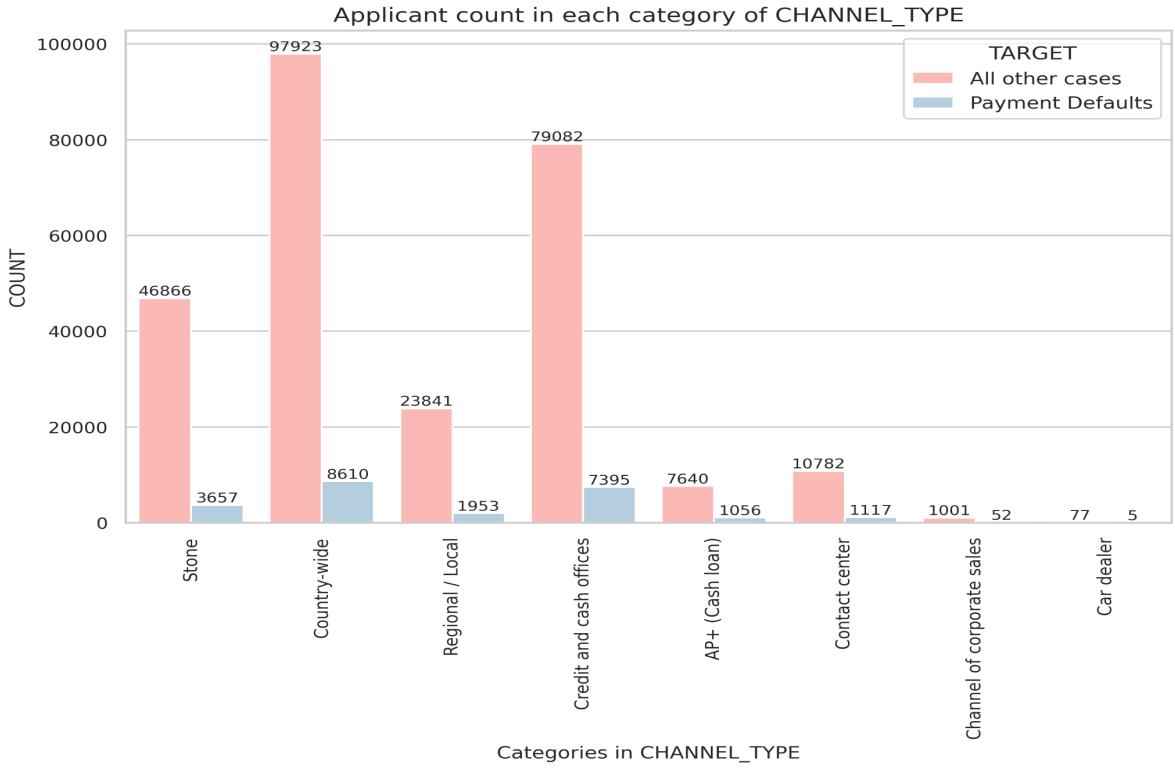
6. **Key Takeaway:** Applicants whose previous loan application was for cars did not have payment difficulties whereas the same for cards have most chances of default.

NAME_PORTFOLIO	Clients with payment difficulties by each portfolio type	All other Cases by each portfolio type
POS	7.57%	92.43%
Cash	8.67%	91.33%
XNA	9.04%	90.96%
Cards	9.70%	90.30%
Cars	6.49%	93.51%



7. **Key Takeaway:** Applicants whose channel of sale for loan was through AP+(Cash Loan) have more percentage of defaults. According to the number of applications and default rate, the best channel for sale is Country-wide.

CHANNEL_TYPE	Clients with payment difficulties by each channel type	All other Cases by each channel type
Country-wide	8.08%	91.92%
Credit and cash offices	8.55%	91.45%
Stone	7.24%	92.76%
Regional / Local	7.57%	92.43%
Contact center	9.39%	90.61%
AP+ (Cash loan)	12.14%	87.86%
Channel of corporate sales	4.94%	95.06%
Car dealer	6.10%	93.90%



Conclusion:-

1. From application_data it can be concluded that according to the number of applications and default rate,
    a. Applicants from the age group 20-30yrs should be avoided.
    b. Although widows have low cases of defaults but due to population size married people have less chances
    c. People without cars should be preferred

    d. Applicants living in house/apartment have less chances of default and who are living with parents or rentals should be avoided.
    e. Applicants with occupation_type Accountants, core staff, high skill tech staff, managers, medicine staff have less chances of defaults.
    f. EXT_SOURCE_3 score and EXT_SOURCE_2 score are good scores to predict defaults.
    g. Applicants with higher education should be targeted for credit sale
2. From merged data it can concluded that according to the number of applications and default rate,
    a. the best channel for sale is Country-wide
    b. Suitable Applicant is that whose previous loan application was for POS.
    c. The best product combination is POS for household with interest
    d. Whose Previous application interest rate are low_normal or middle

Result:-

I have answered all the questions asked by the company in this project and explained the result and conclusion under the project insights part. While doing the project I applied my learning of statistics and understanding of different functions, basics of programming, Python programming, visualization packages including matplotlib, seaborn, plotly, dataframe analysis packages including pandas, numpy used in python.