CREDIT EDA CASE STUDY

-KSHAMA SHETYE (DSC31)
-RAUNAK BASU (DSC31)

TABLE OF CONTENTS

INTRODUCTION - PROBLEM STATEMENT & GOALS

→ When a client applies for loan, the company has to decide for loan approval based on the applicant's profile.

→ Two types of risks are associated with the company's decision:

- If the applicant is likely to repay the loan (Non- Defaulter), then not approving the loan results in a loss of business to the company.
- If the applicant is not likely to repay the loan(Defaulter), then approving the loan to such a client will lead to a financial loss for the company.

→ **Our goal** is to analyze the new and previous application data to study the pattern of defaulters (TARGET) and give recommendations based on the observations which should help the company to make informed decision based on past data, and reduce the number of defaulters in future.

EDA METHODOLOGY

Step 01

Cleanup and Analysis of New Application Data

Step 02

Cleanup and Analysis of Previous customers Application Data

Step 03

Overall Analysis and observations on the combined All Application data

Step 04

Analysis of results for Recommendations

DATA CLEANING

- Before any data is used, it needs to be cleaned and treated to get the best possible views for analysis.
- Following steps were performed on each data set.

Understanding the Structure

The structure of the data is observed and unwanted (Data with nulls) or data not important for analysis are removed

Datatype Adjustments

The datatype of columns that are misrepresented are corrected

Conversions

Data is converted to usable formats or converted to other formats as needed(days to years) when needed

Outlier and Imbalance Detection

Outliers and Imbalance in the data are detected and reported for imputation.

Imputation

Imputation is done on the data as per column by choosing values to impute based on the outliers

Plotting and analysis is done after cleaning the data.

Suitable scale and further imputation is chosen for very specific cases where removing the data may cause loss of valuable important information

UNDERSTANDING THE DATA

NEW APPLICATIONS

Contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

PREVIOUS APPLICATIONS

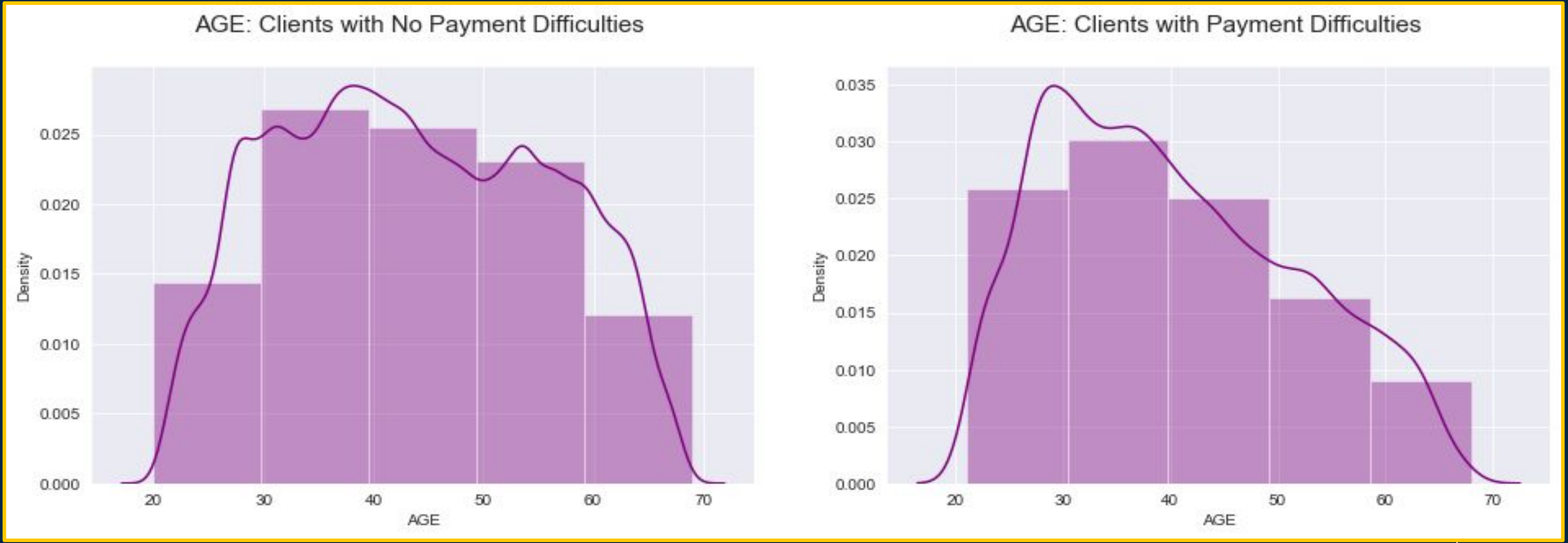
Contains Information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

ALL APPLICATIONS

ANALYSIS AND INFERENCES

ANALYSIS OF THE NEW APPLICATION DATA

AGE DISTRIBUTION OF THE APPLICANTS



- Majority of the clients with payment difficulties are between the age of 30-40, and then it keeps on decreasing.
- Most of the clients with no payment difficulties are in the age group 30-60.

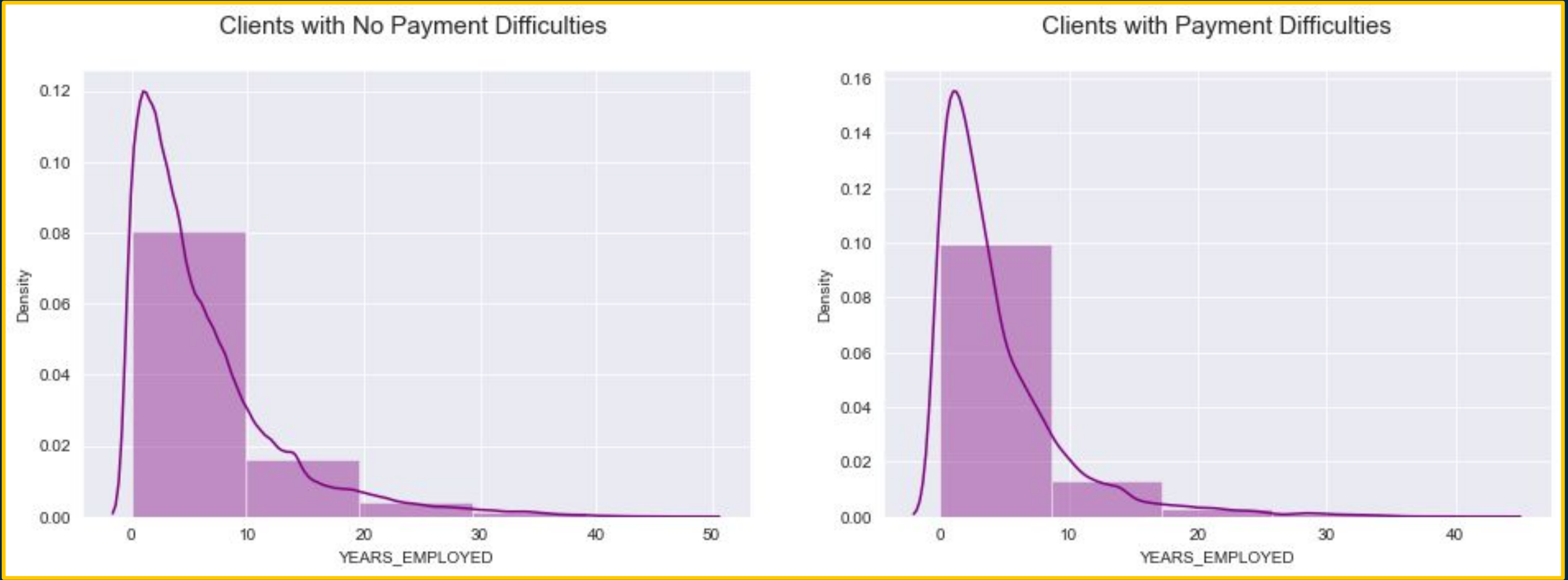
YEARS EMPLOYED

Inference -

- For both clients with or without Payment difficulties, the peak number of applicants are working with the current organization for less than 5 years of employment and drops considerably after that.

Note -

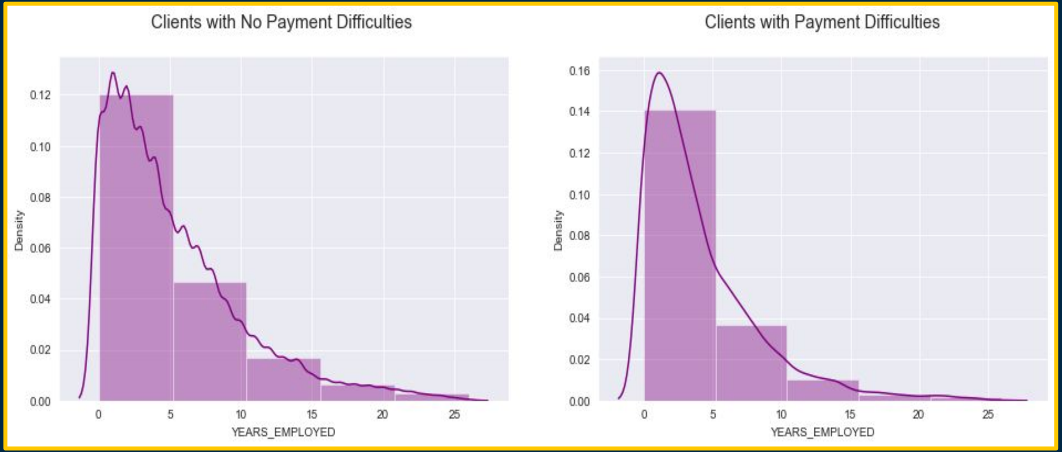
- The data regarding Days Employed of the client has outliers which were recognized in our analysis and predictably skews the data, the plots were corrected to remove such outliers by adding a reasonable limit to the employment years.



SEGMENTED ANALYSIS OF APPLICATION DATA

- It is apparent from our previous analysis that a big portion of the applicants are between the ages of 25-45 years.

- Segmented analysis aims to provide a deeper insight into this age group.



YEARS EMPLOYED

- Segmentation of the data further emphasizes that the large number of applicants are within 0-5 years of employment in the current organization.

AGE

- The age distribution shows higher number of defaulters at the lower end of the age group.

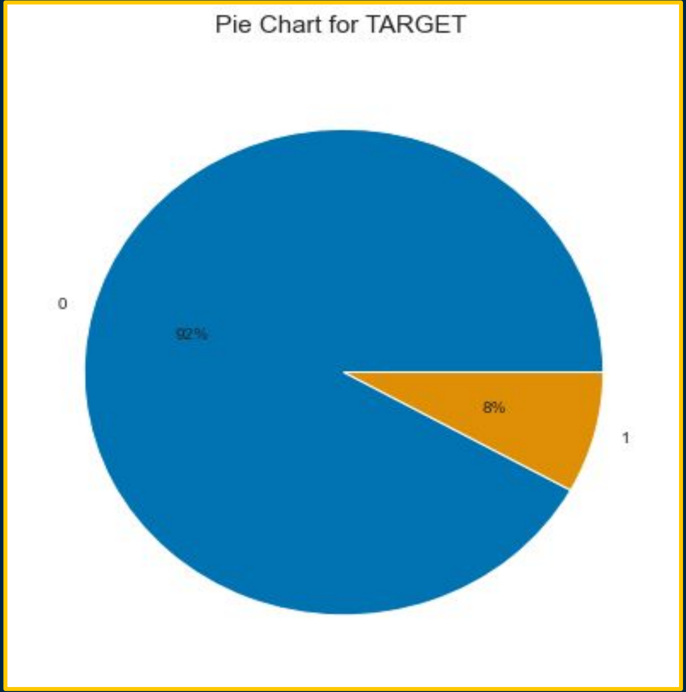
IMBALANCE IN DATA

- As we can see that, there is high data imbalance between clients with payment difficulties and the other.

- Approx 92% of clients have paid the loan on time, while approx 8% of clients faced difficulties in paying the loan on time.

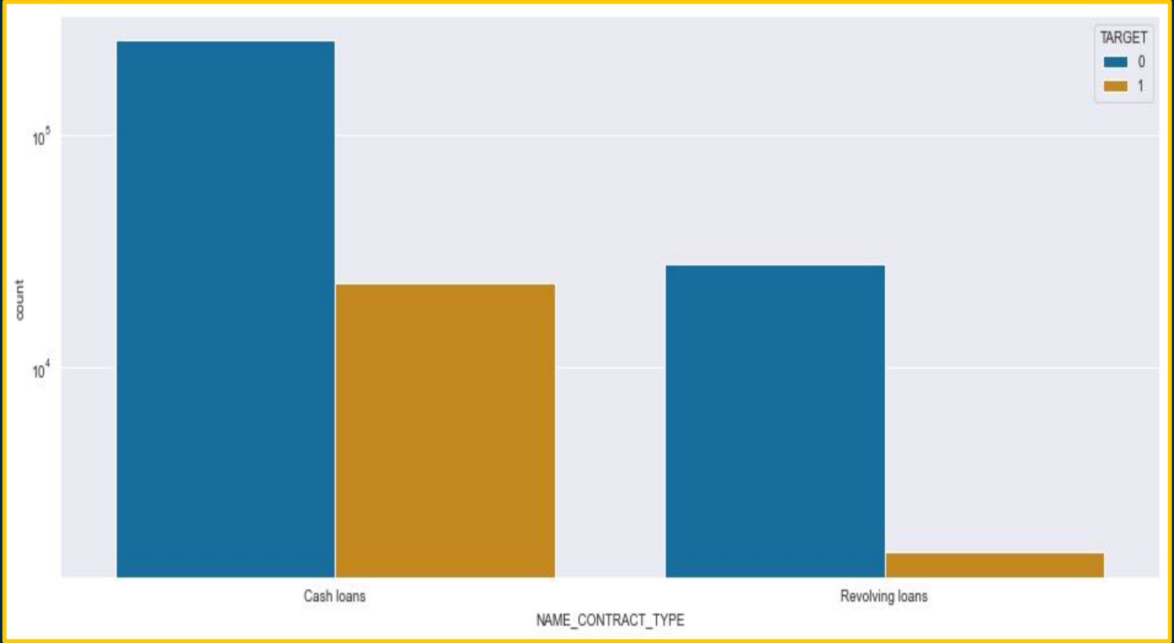
- The ratio of data imbalance for TARGET variable is 91.92 : 8.07

- The data is divided into two parts, Clients with Payment Difficulties (TARGET 1) and Clients with No Payment Difficulties (TARGET = 0) for further analysis.



CONTRACT TYPE

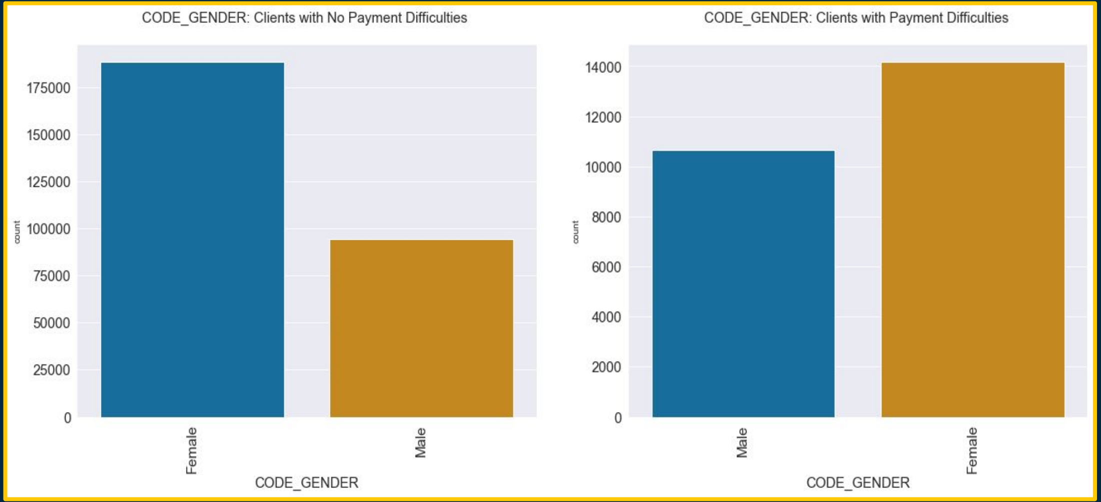
0 = Client with No Payment Difficulties, 1= Client with Payment Difficulties

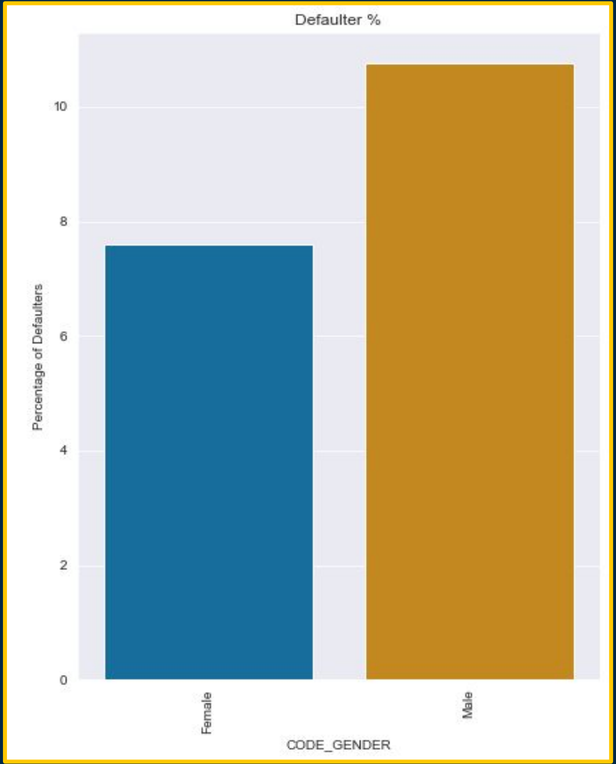


This graph is plotted on the log scale to improve readability

- Cash loans overall are more popular than the revolving loans.
- While revolving loans have a comparatively higher number of defaulters

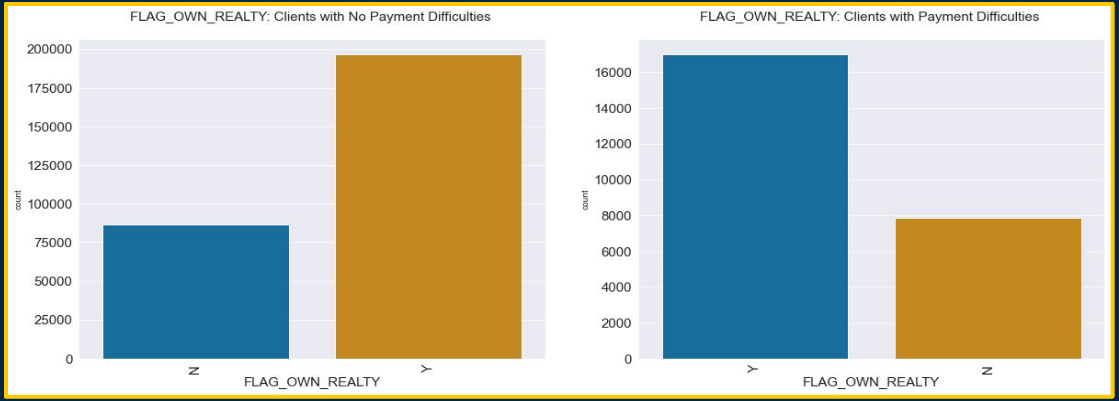
GENDER

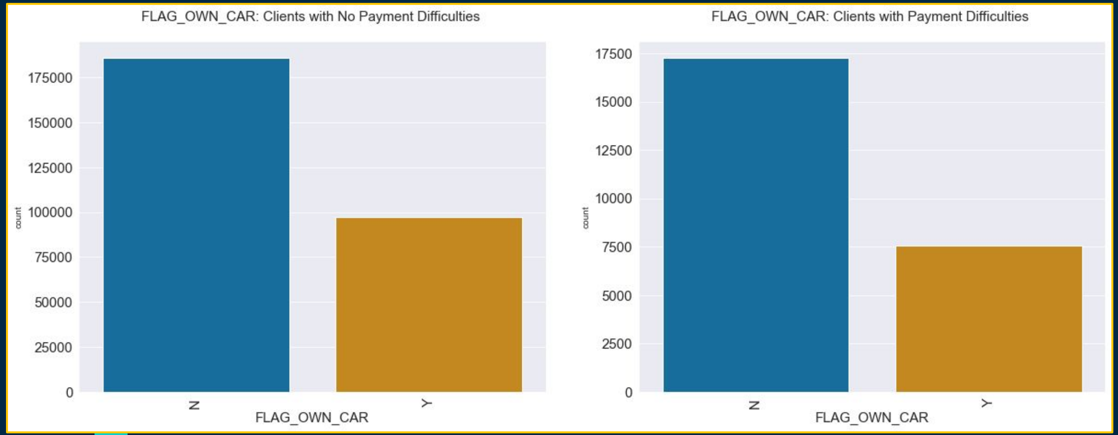




- Males have defaulted more than females, despite there being more female clients.

ASSETS(CARS / REALTY)

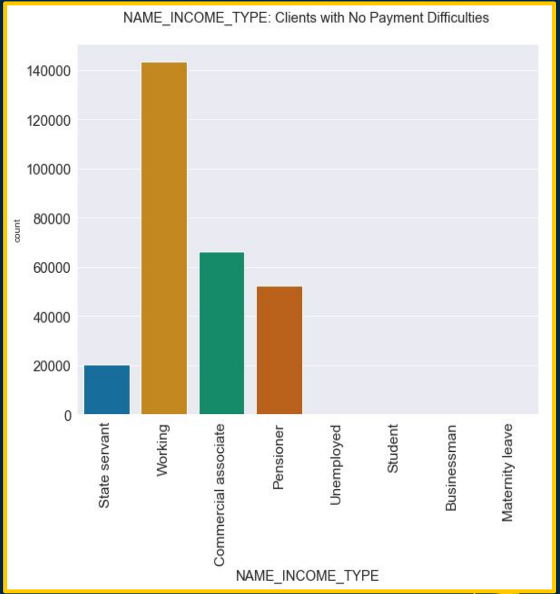




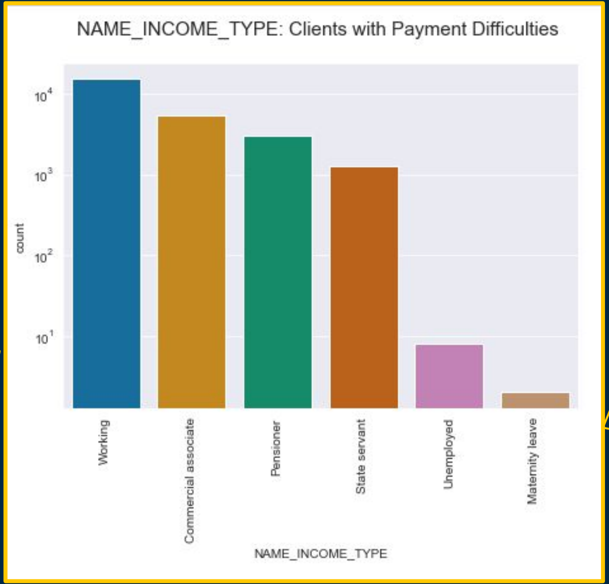
- Considerable number of clients with payment difficulties have realty, which might be the indicator of more ongoing loans and bigger expenses.

- Same cannot be said about the client owning a car, as those without cars, have defaulted more than those who have cars.

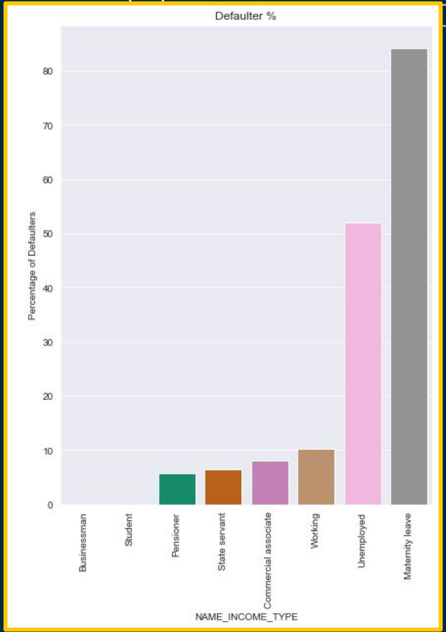
INCOME TYPE



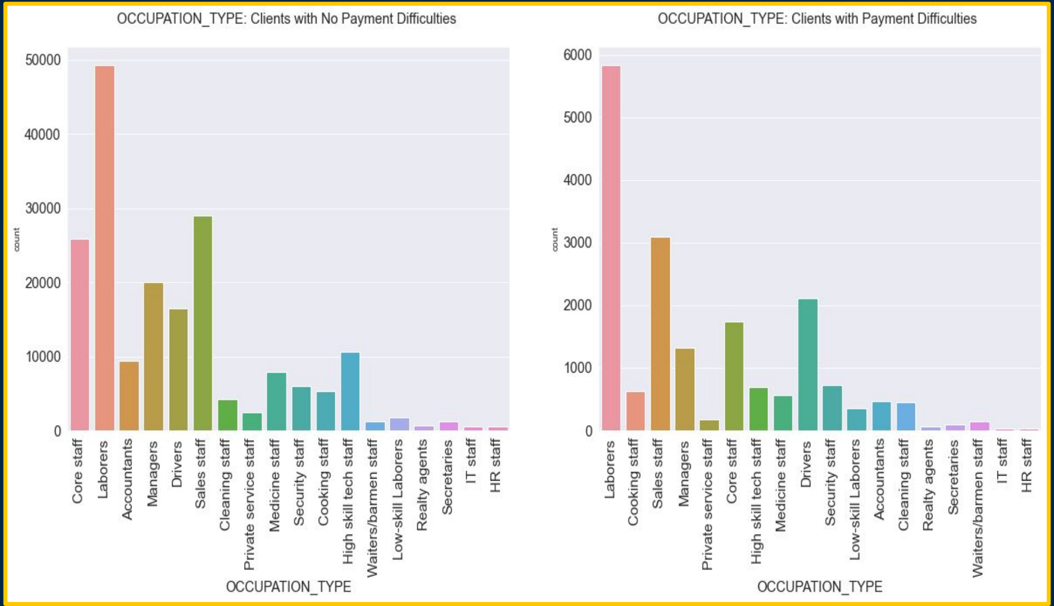
- The defaulter percent is high in **maternity leave** and **unemployed category.**
- While the count of non-defaulters is high in working income type. Followed by commercials and pensioners.

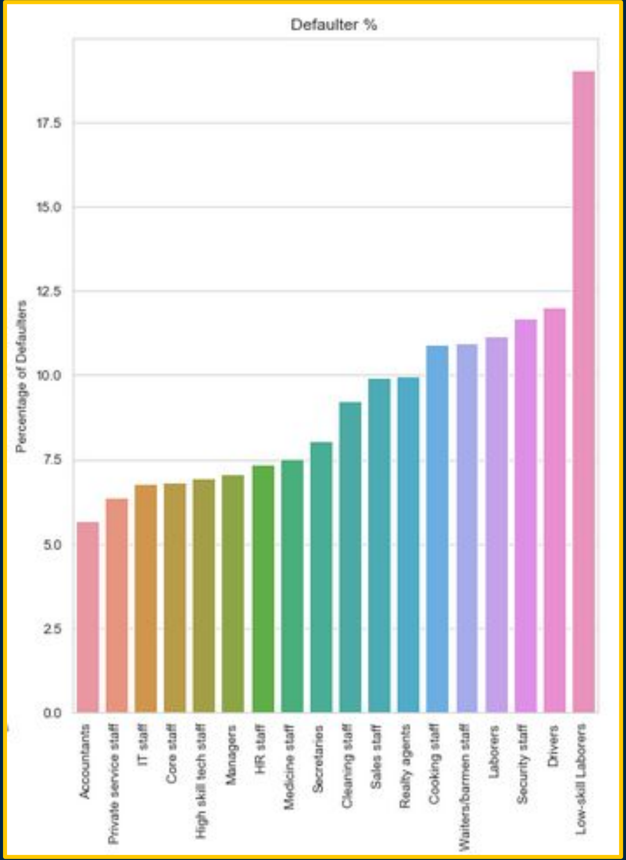


Logarithmic View of the Defaulter's Income type



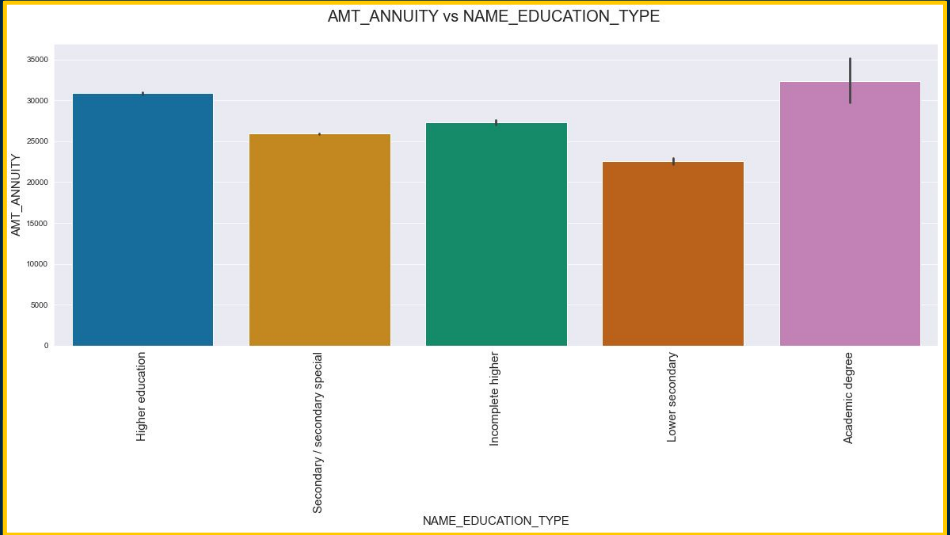
CLIENT OCCUPATION

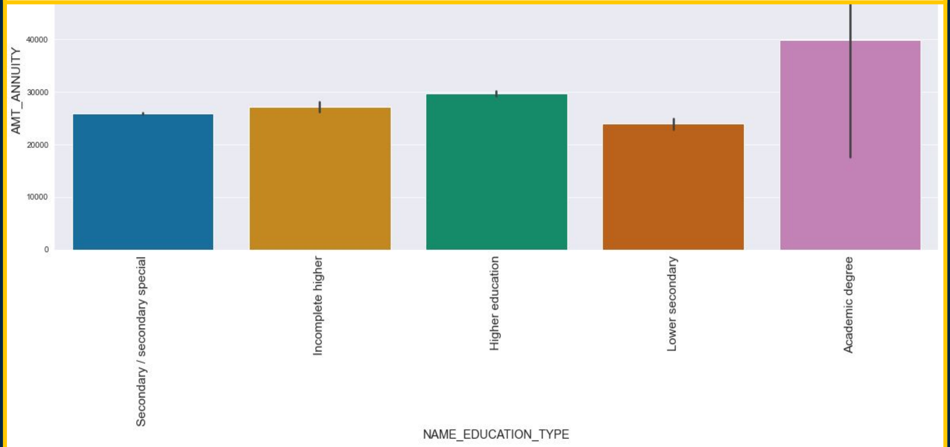




- A big portion of the clientele is from occupation of Laborer, Sales Staff and core staff.
- Low skill laborers are by far the biggest offenders, followed by Drivers and Security staff.

ANNUITY VS EDUCATION

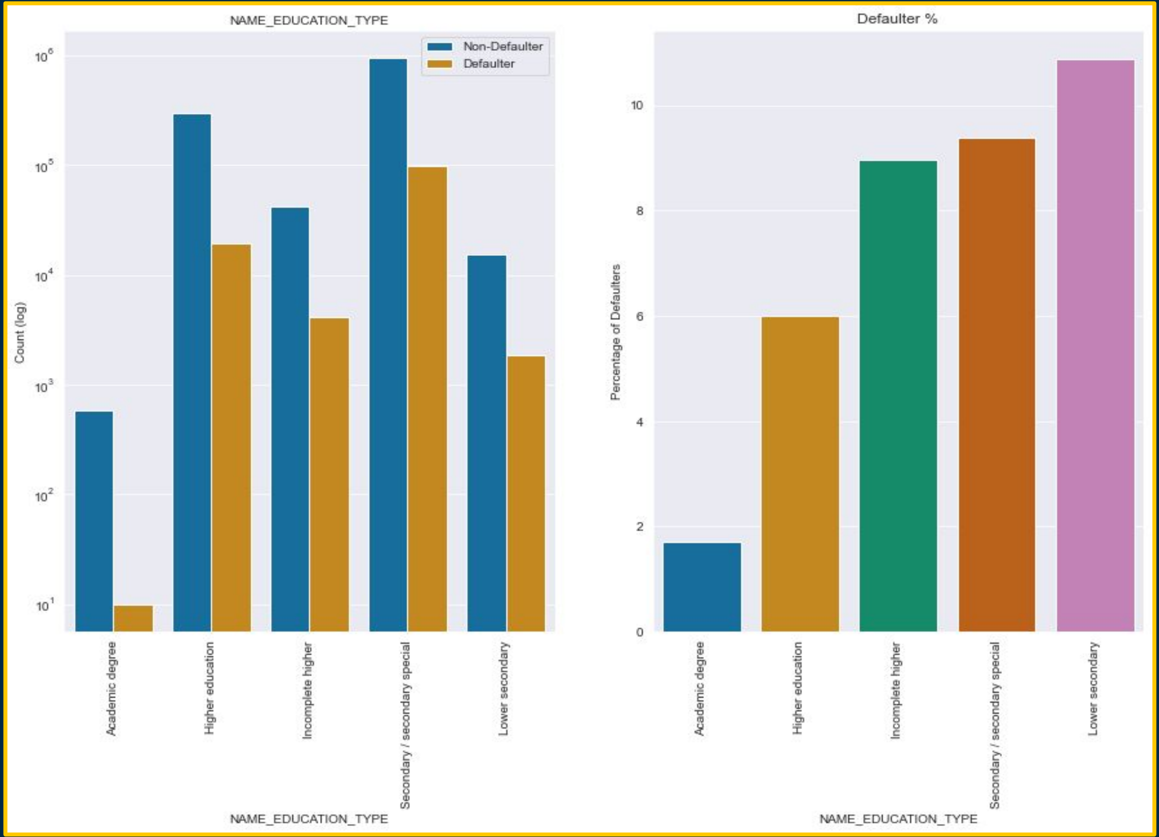




- **<u>Clients without Payment Difficulty</u>**
  People with higher academic degrees are more likely to pay a much higher annuity, followed by clients with higher education

- **<u>Clients with Payment Difficulty</u>**
  Clients with academic degrees and with around 40,000 in annuity seem to have problems.

DEFAULTERS BY EDUCATION STATUS



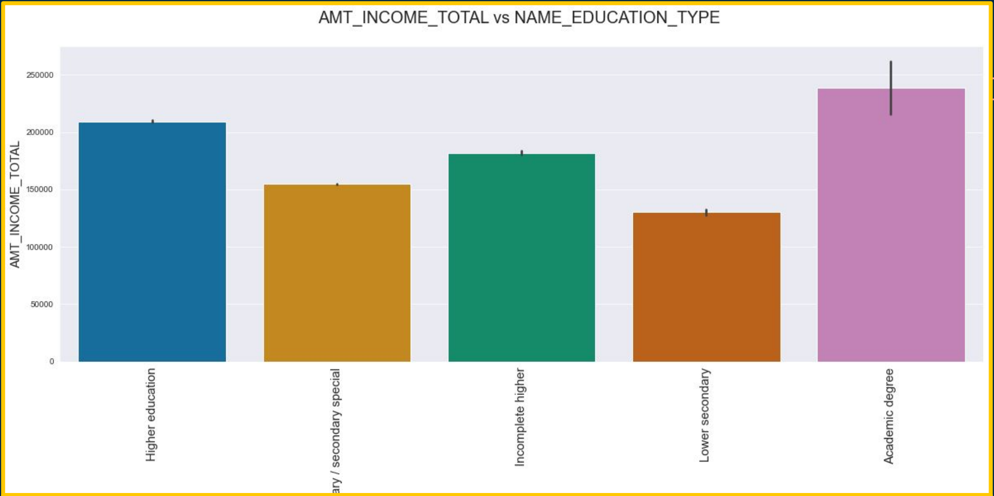
- Clients having **Lower Secondary education** have the highest percentage of default followed by **secondary education.**

- Education is an important factor while approving loans as academic degree holder clients have quite low % of defaults.

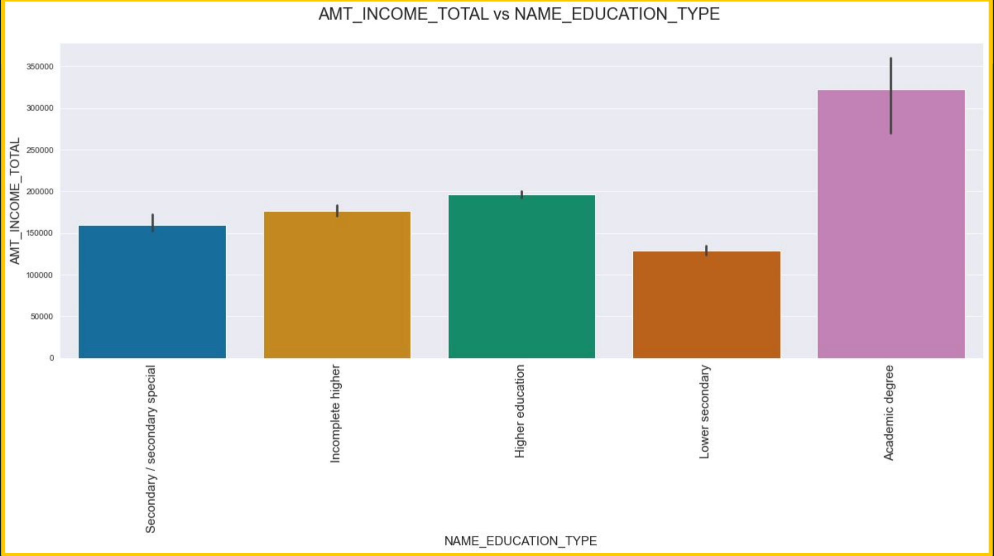
EDUCATION VS INCOME

- Clients with academic degrees have higher overall income.
- For Clients with no payment difficulties, income seems to be around 250k, while for defaulters it is much higher at over 300k.
- This might indicate a need to adjust the credit limit dispensed to such clients with higher incomes.

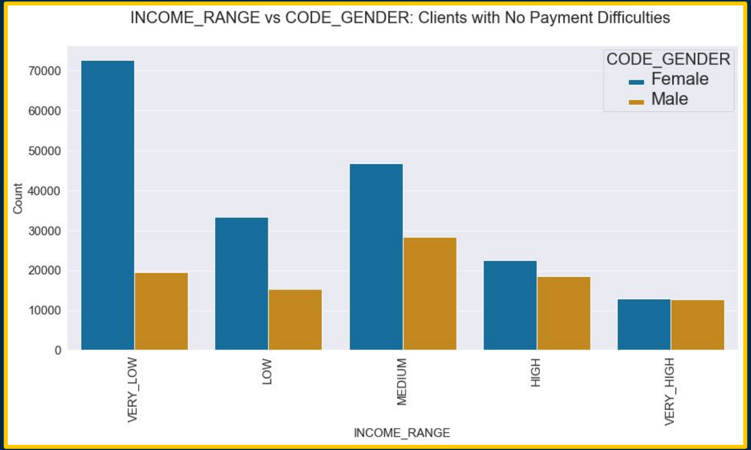
Clients with no Payment Difficulties

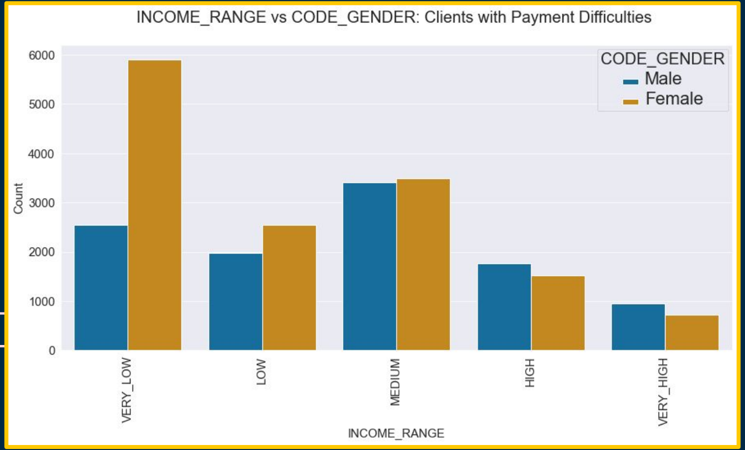


Clients with Payment Difficulties



INCOME RANGE VERSUS GENDER

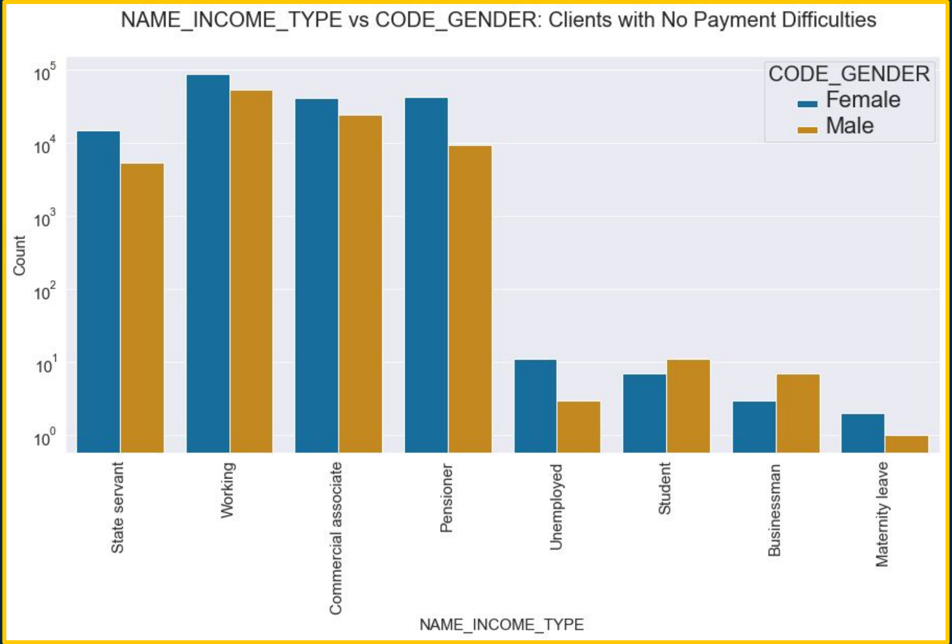


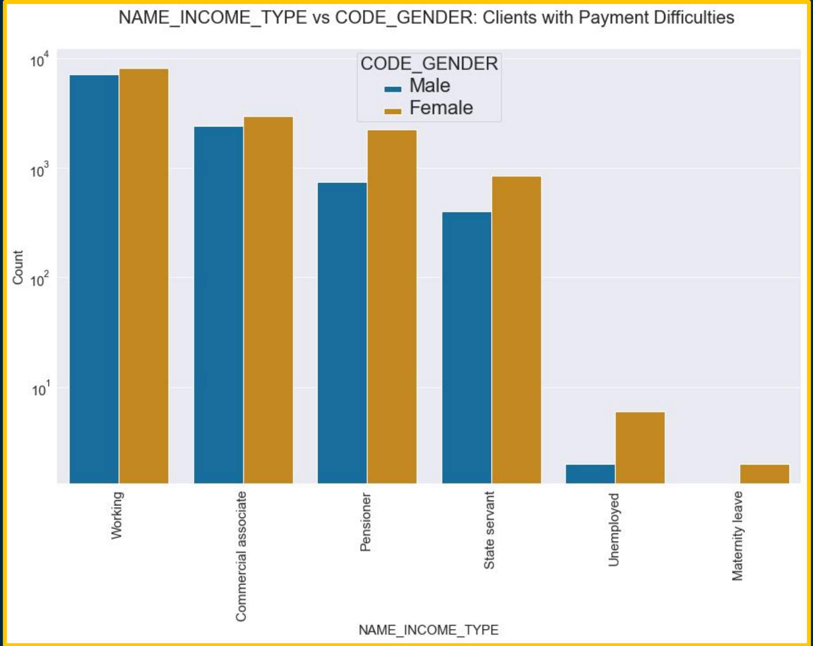


- **Clients with No Payment Difficulties-** The ratio of females to males in the very_low income group is very high, while females are more in number across all income groups (High count of female clients in data) the ratio in every other group is relatively closer.

- **Clients with Payment Difficulties-** The number of defaulters who are males are higher than the females in the high and very_high income groups

DISTRIBUTION OF CLIENT BY INCOME TYPE AND GENDER

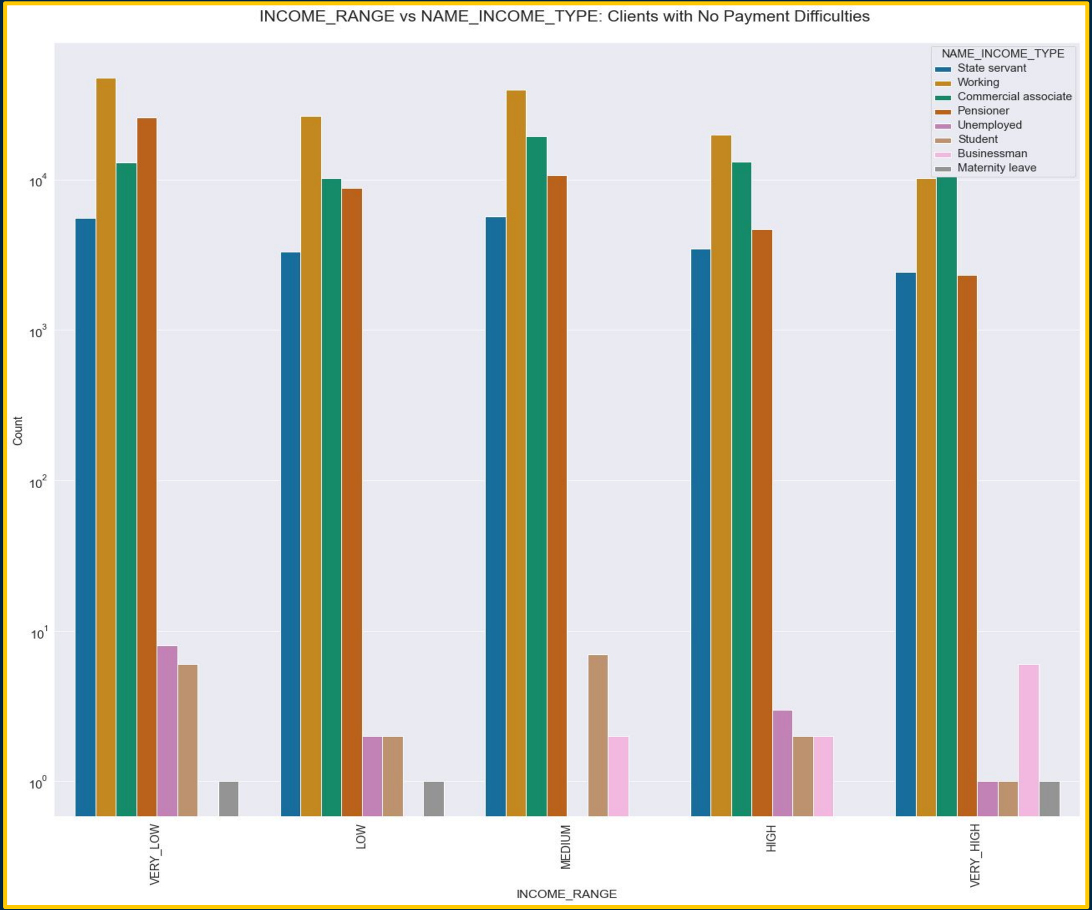




- For Clients with No Payment Difficulties, we can see that for most of the income types, the females are high as compared to males. Except for students and businessman where we have more male clients.

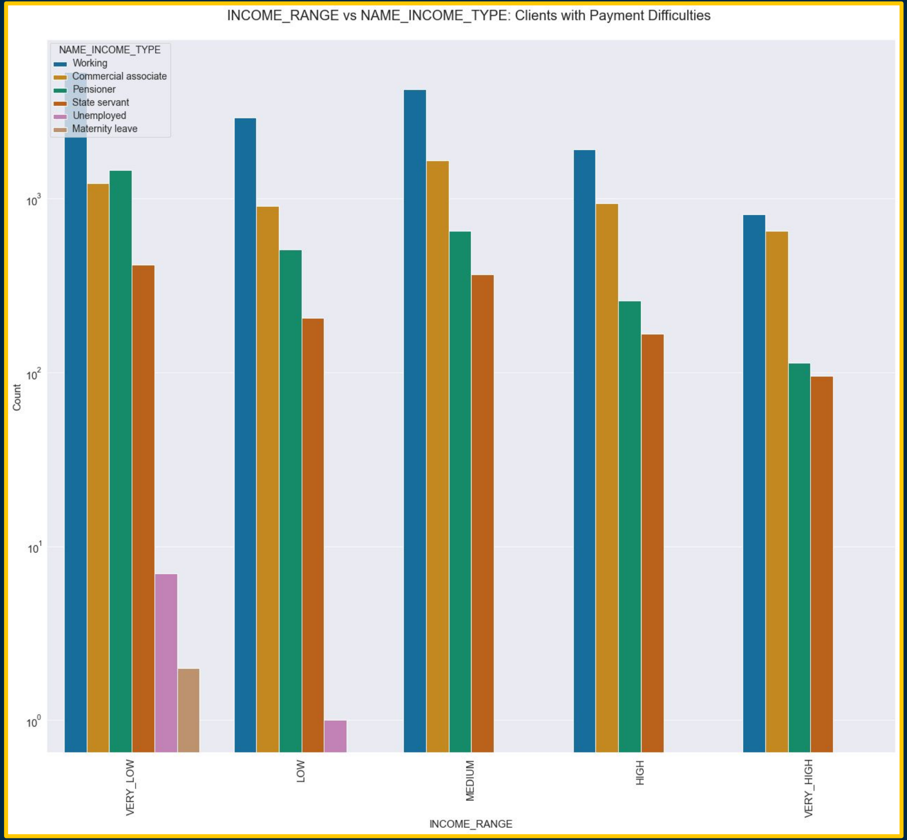
- For defaulters, students and businessman have negligible defaulters in the data.

DISTRIBUTION OF CLIENT BY INCOME TYPE AND INCOME RANGE (Part-1)



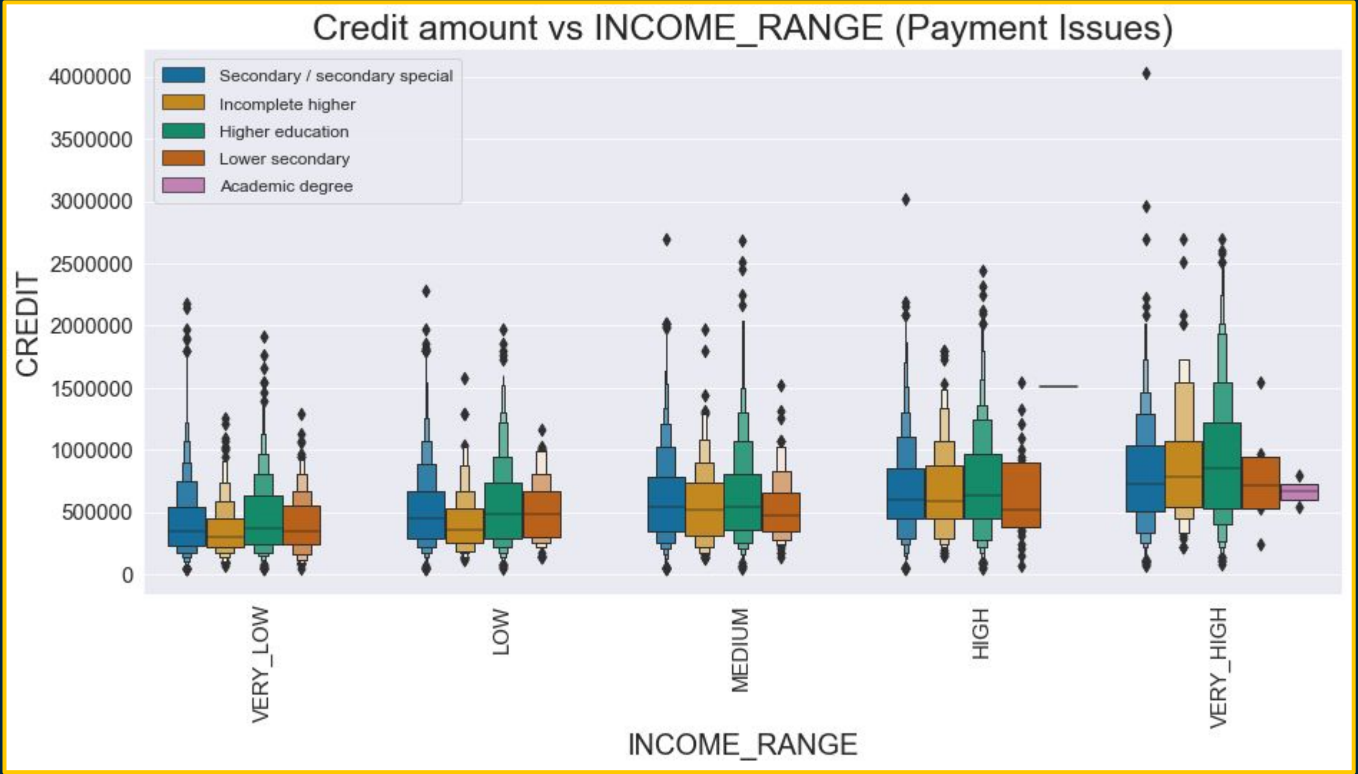
- **Clients with No Payment Difficulties**, the working and commercial associates, have an equal count in the very_high income_range. The company should focus more on these clients for business.
- While in the case of very_low income_range, the working income_type has the highest count, followed by pensioners and then the commercial associates.
- The income_range count of state servant is very close in the category of very_low and medium, followed by high and low. The company can approach these customers due to their stable income flow, but only after thorough investigation.

DISTRIBUTION OF CLIENT BY INCOME TYPE AND INCOME RANGE (Part-2)



- Clients with Payment Difficulties, the working count is slightly more than commercial associates count in the very_high income range.
- While in the case of very_low income_range, the working income type has the highest count, followed by pensioners and then the commercial associates.
- The unemployed are seen only for very_low and low income_range.

CREDIT AMOUNT vs INCOME RANGE (Clients with Payment Issues)



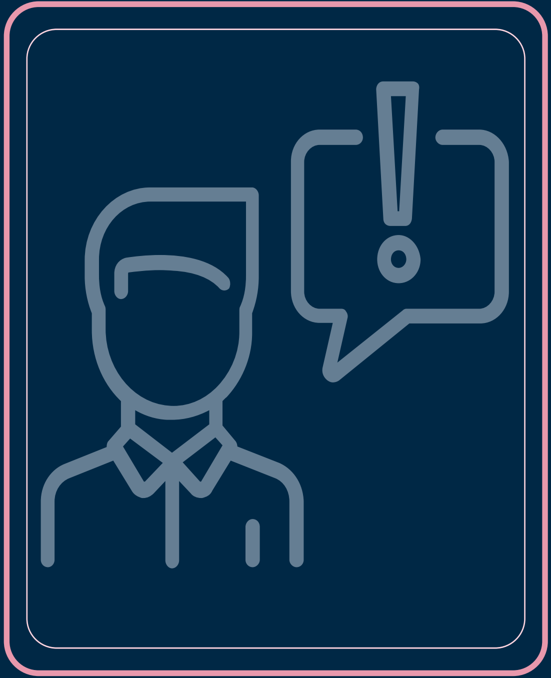
- For clients facing payment problems, it is usually with people with higher education or with secondary education who have taken the biggest credits.
- Clients with academic degrees are far less likely to default across all income groups.

ANALYSIS ON THE PREVIOUS APPLICATION DATA

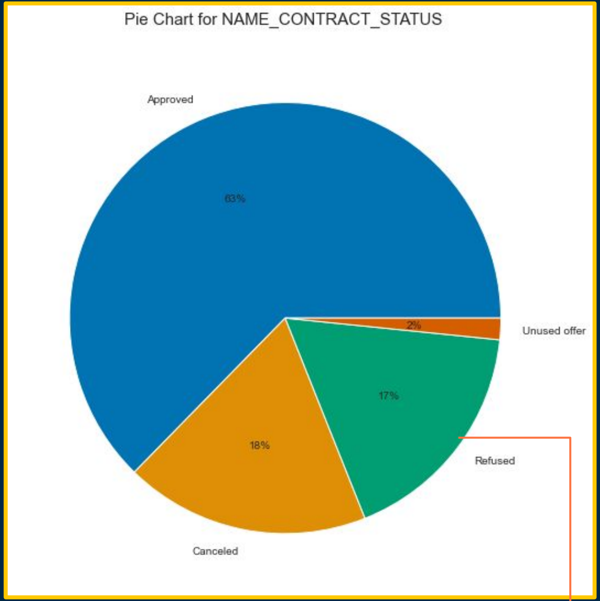
WHAT IS THE PREVIOUS APPLICATION DATA

- Previous Application contains information about the client's previous loan data.

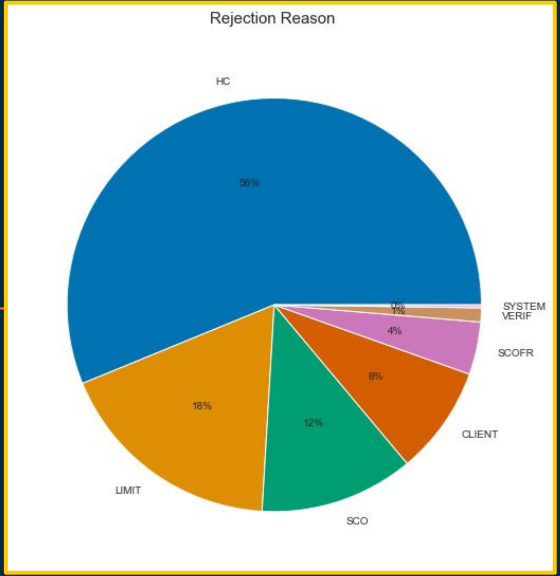
- It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.



PREVIOUS CONTRACTS

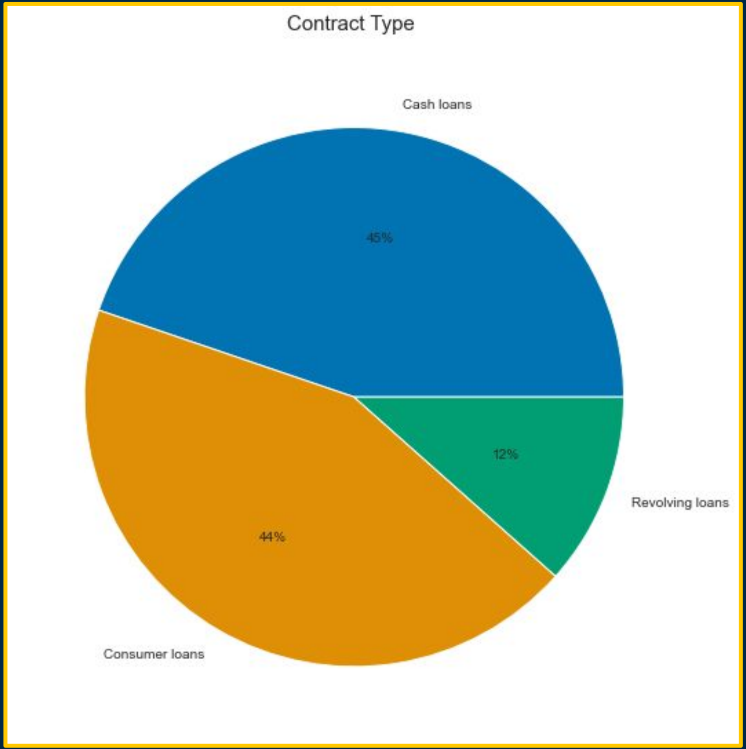


- **63%** of the previous applications were approved.
- Only **17%** of the applications were **refused.**



- **56%** of the rejections were due to HC
- LIMIT and SCO together accounted for 30% of rejections

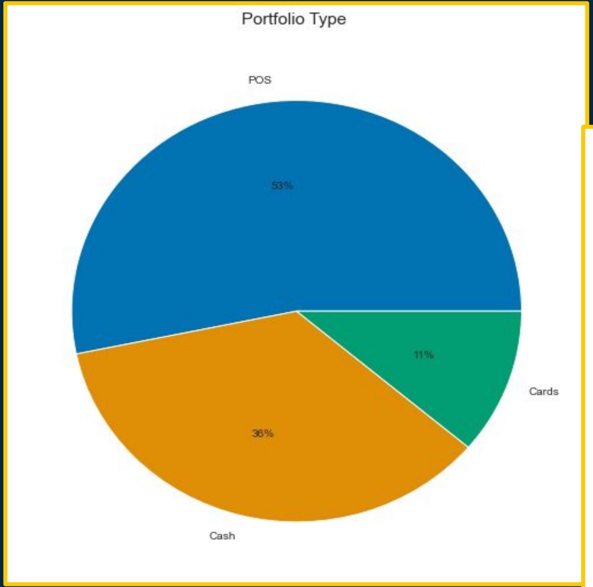
CONTRACT TYPES

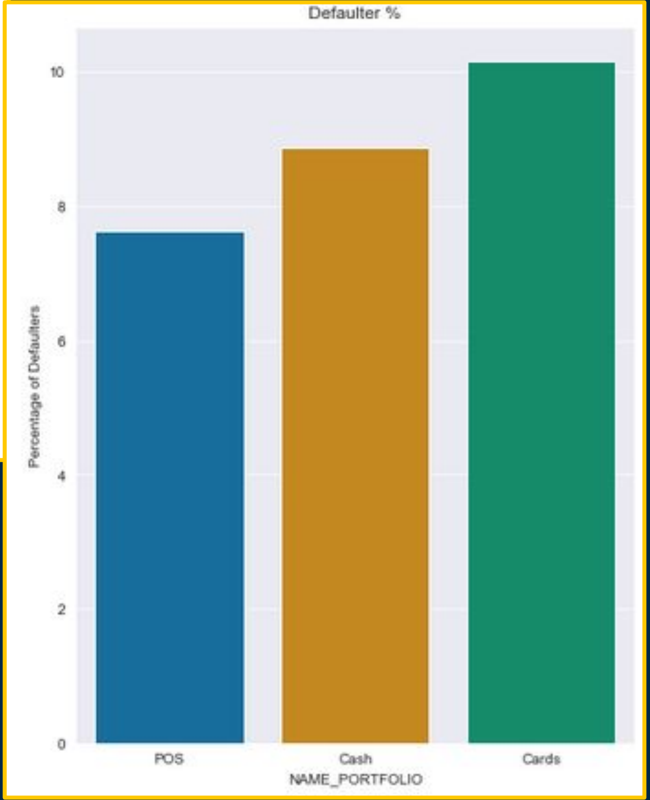


- **Cash** and **Consumer** loans were the most popular contract types in all the previous applications, accounting for over **89%** of all past applications.

- Only **12%** of clients in the past applied for revolving loans

PORTFOLIO TYPES

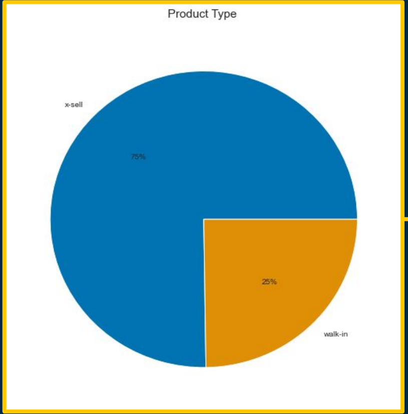


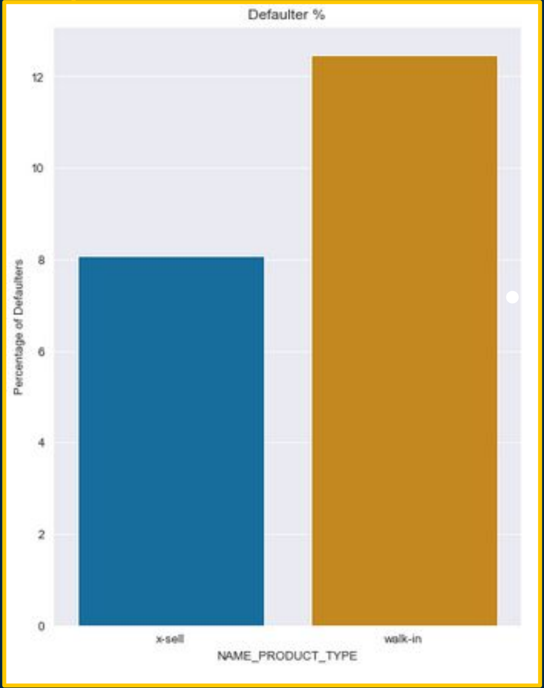


- **53%** of the past applications were for **Point-of-Sale (POS)** credit, followed by **Cash** at **36%.**

- PoS has the least percentage of defaulters.

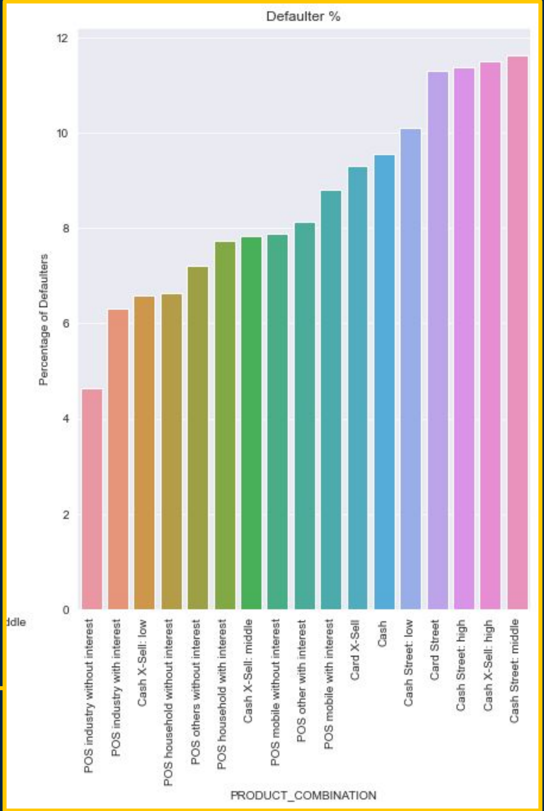
PRODUCT TYPE & COMBINATIONS (Part 1)



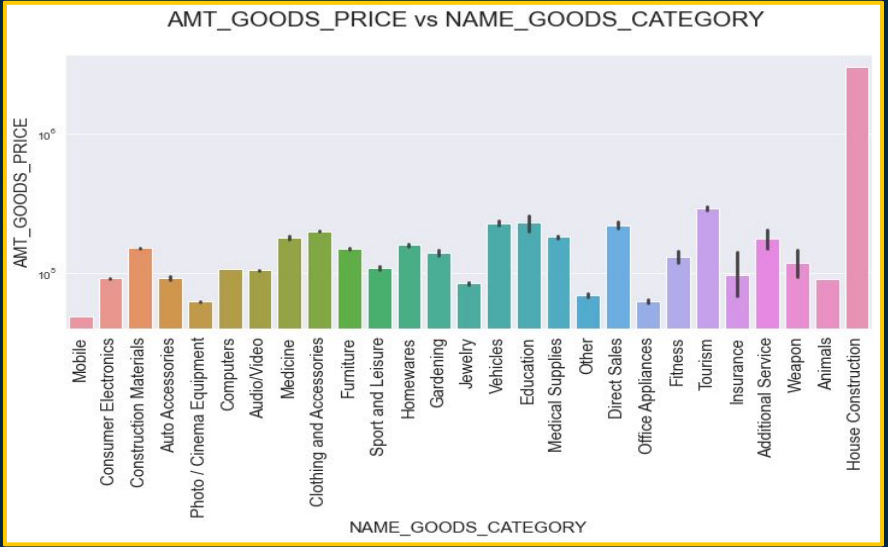


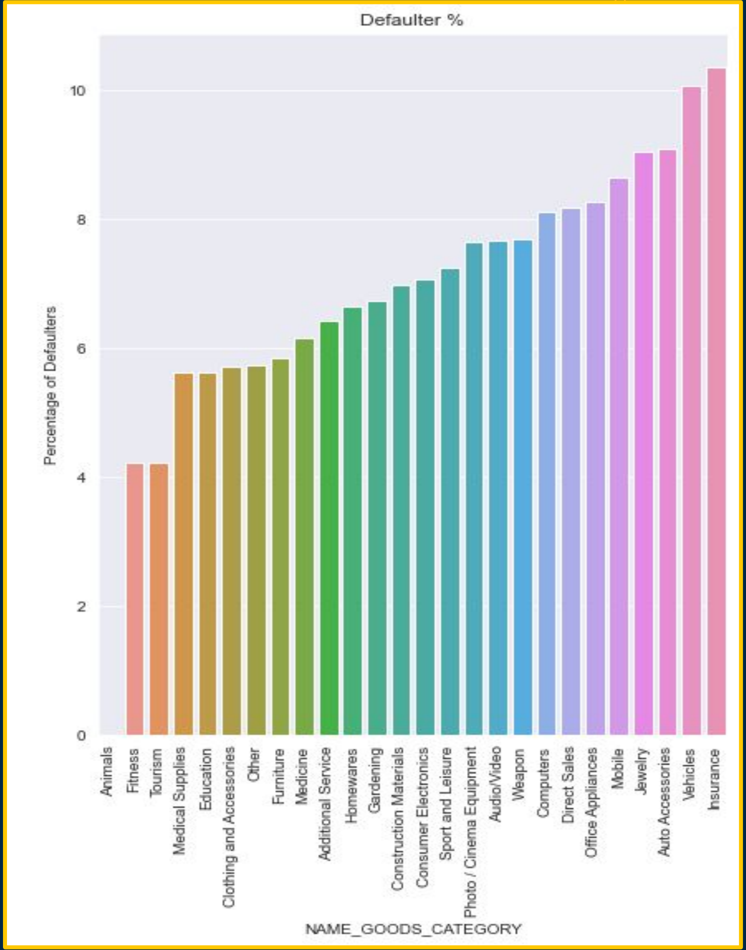
- Cross-Sell (x-sell) accounted for the majority of the previous applications at **75%**. While rest were walk-ins.
- Despite accounting for the 75% of Previous applications, **x-Sell** has a significantly lower default rate than walk-in clients.

In case of product combinations Card Street,**Cash Street(high & mid)**, **Cash x-sell** are all high risk products with high default %.



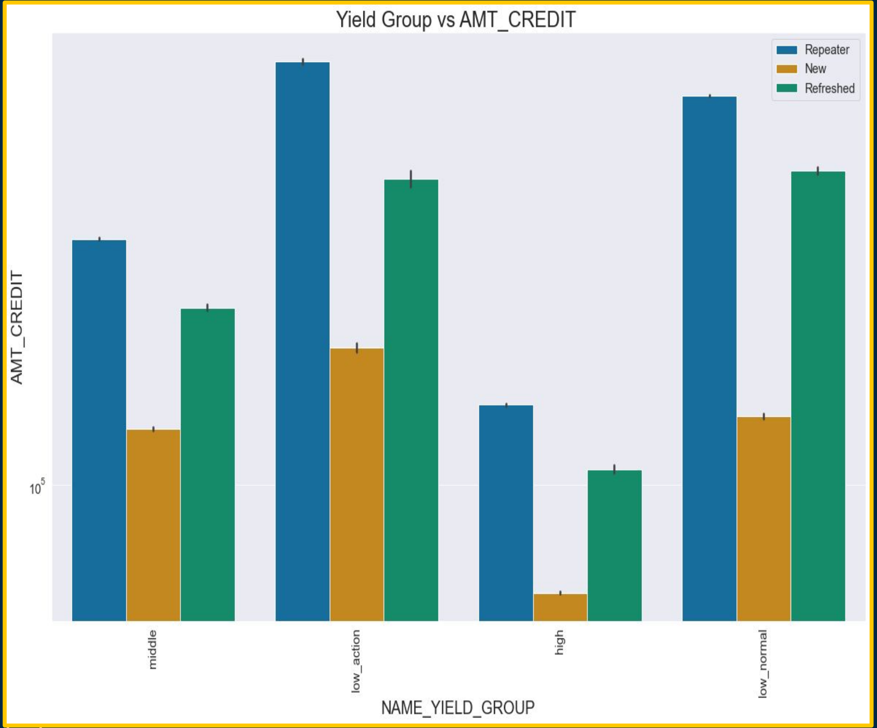
GOODS PRICE vs CATEGORY

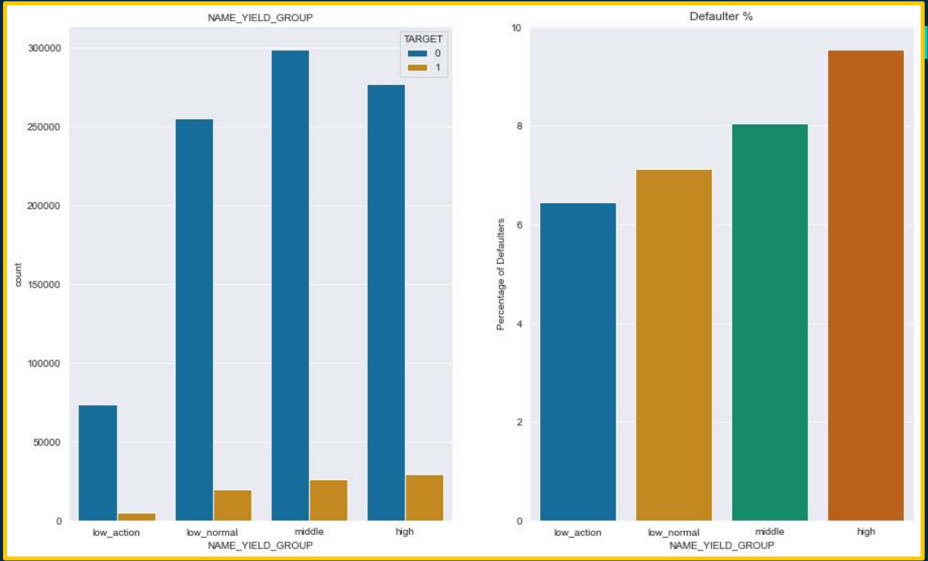




- House construction has been the most expensive applications in the past. Followed by Tourism, Education and Vehicles.
- However, the most defaults are found in categories of Insurance and Vehicles

CLIENTS IN YIELD GROUPS

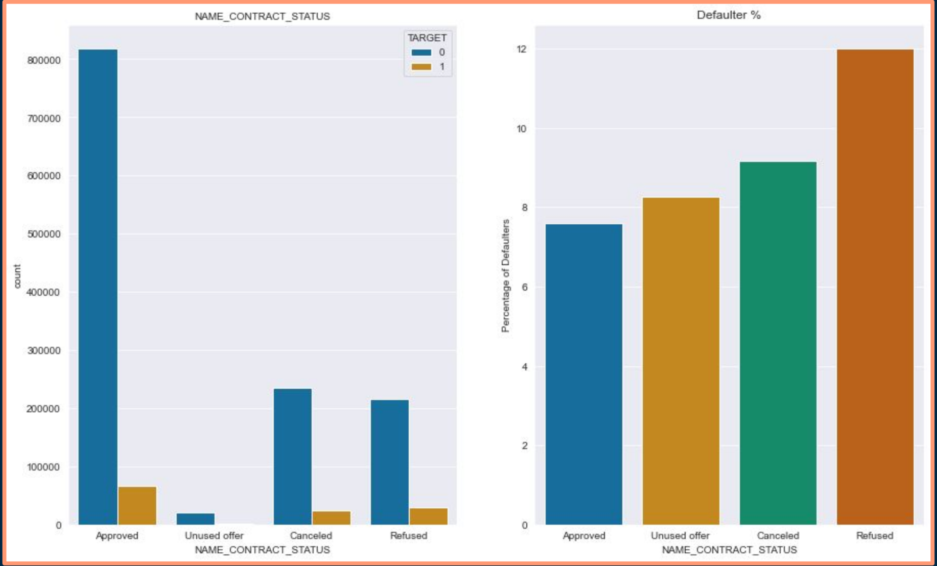




- High yield group has the highest percentage of defaults where as low_action group has lowest among them.
- Also, there are many clients who pay on time with high yield.

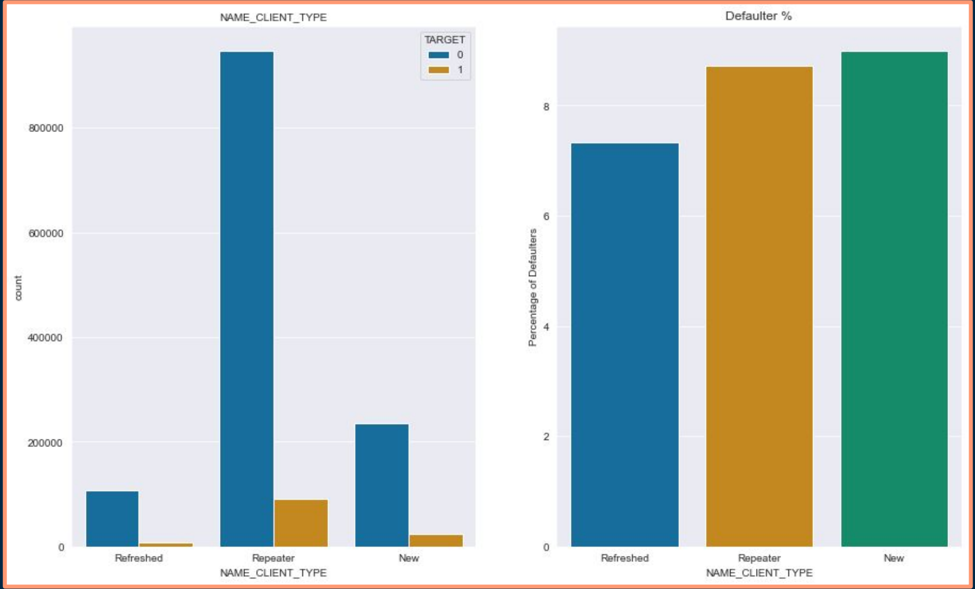
ANALYSIS ON THE COMBINED DATA (ALL APPLICATION)

CLIENT's HISTORY

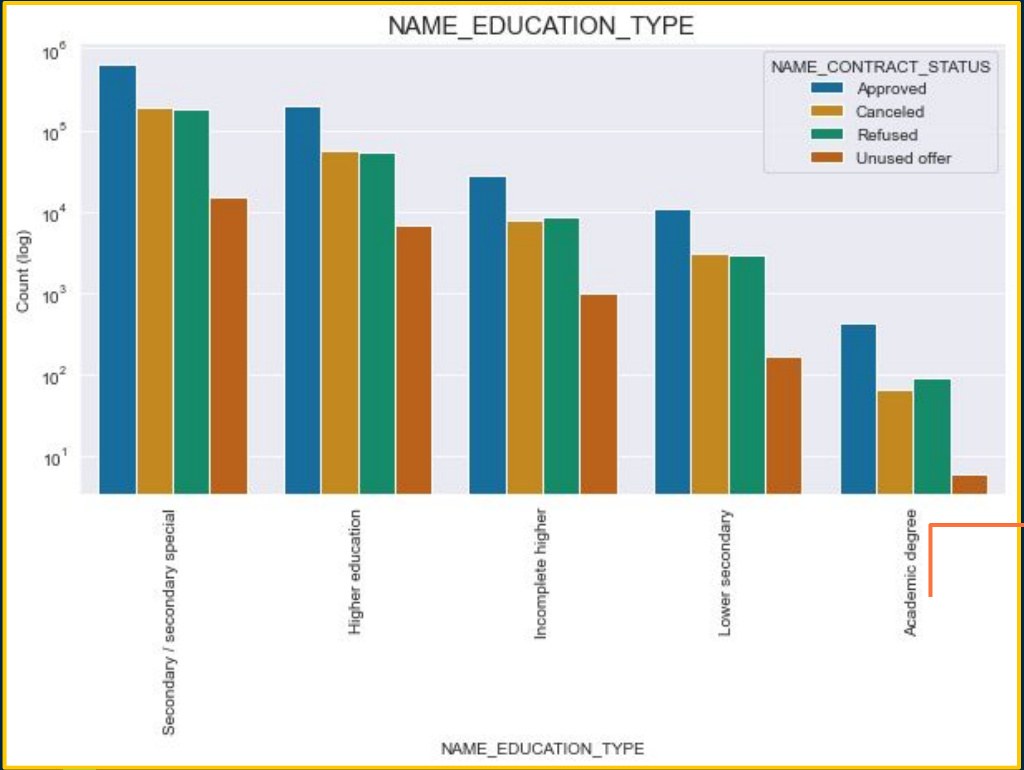


Clients who were previously refused, have a much higher rate of default.

The chance of newer clients to default is slightly more than the repeated or refreshed clients.



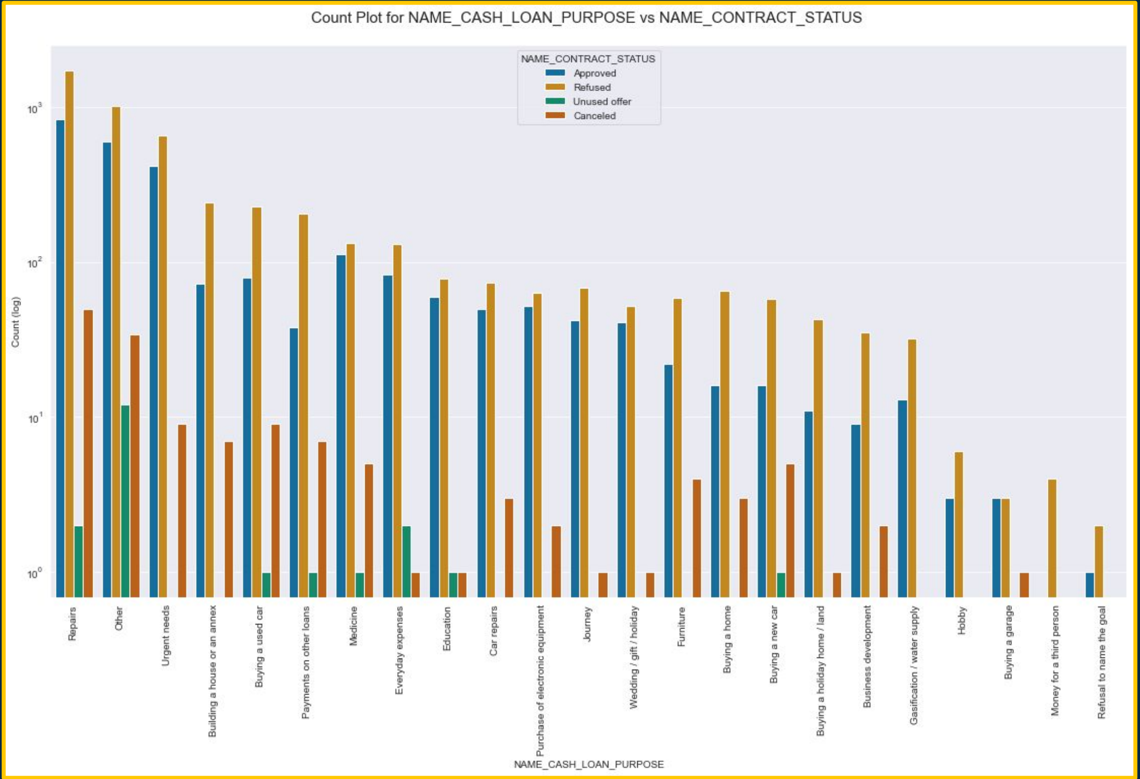
CLIENT's WITH ACADEMICS



- Clients with Academic Degrees have the highest approval rate and lowest refusal rate among all the education groups

Academic degree	Approved	418	72.19%
	Refused	91	15.72%
	Canceled	64	11.05%
	Unused offer	6	1.04%

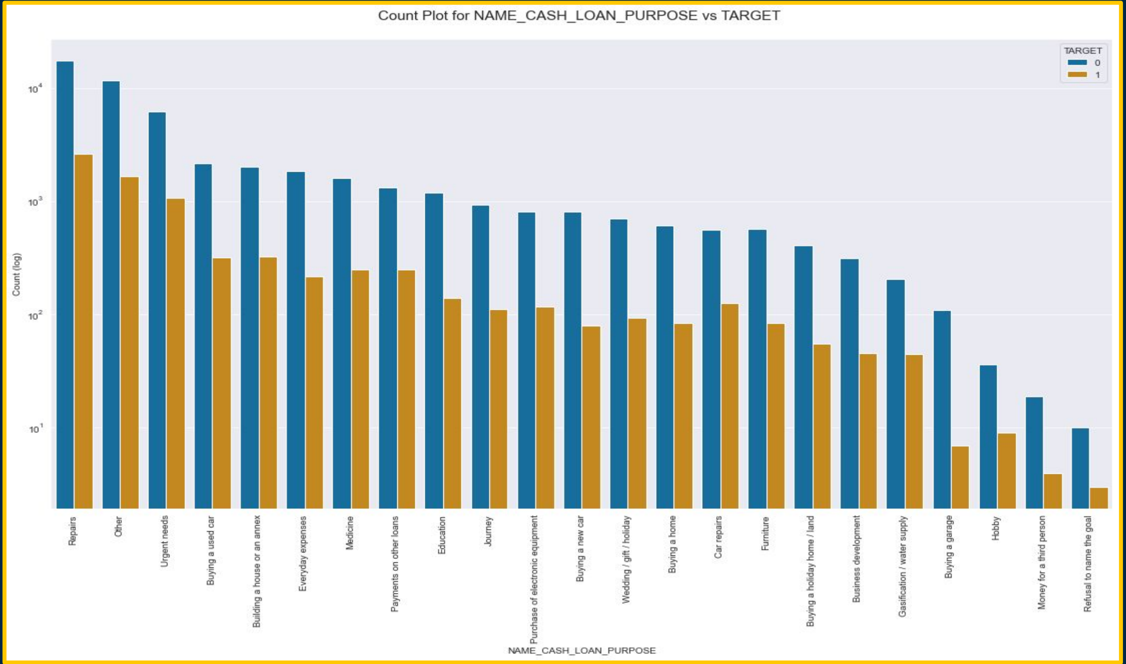
LOAN PURPOSE AND STATUS



- Loans that were taken for **Repairs** and **Other** have high refusals and rejections by both clients and the company.

Approved	Repairs	7377	31.97%
	Other	5744	24.89%
	Urgent needs	3105	13.46%
Refused	Repairs	12176	35.0%
	Other	7354	21.14%
	Urgent needs	4016	11.55%
Canceled	Repairs	527	38.5%
	Other	246	17.97%
	Urgent needs	115	8.4%
	Building a house or an annex	86	6.28%
Unused offer	Other	88	48.09%
	Repairs	37	20.22%
	Everyday expenses	20	10.93%

LOAN PURPOSE AND DEFAULT



- Loans that were taken for **Repairs** and **Other** have the maximum number of defaults.

1	Repairs	2616	33.8%
	Other	1673	21.62%

OBSERVATIONS

Men are more likely to default than females

Clients with better education background are far less likely to default

Clients who were previously refused, have a much higher rate of default.

People who were employed for less than 5 years at their current job have defaulted the most.

Unemployed and those on Maternity Leave are the worst offenders

Point of Sale (PoS) portfolios have the least number of defaulters

People who own realty are more likely to default

Cross-Selling overall is safe, but some combinations of cross-selling have very high risk(Cash Street(high & mid), Cash x-sell)

Credits given for Insurance and Vehicles have very high default

OBSERVATIONS

Refreshed clients and repeated clients are safer and have low risk of default than newer applicants

Loans that were taken for Repairs or **Other** were most rejected by both company and client. These loans also had the highest number of defaults

High yield group has the highest percentage of defaults,whereas low_action group has the lowest among them

Clients with Academic Degrees have the highest approval rate and lowest refusal rate among all the education groups

RECOMMENDATIONS

- Payment difficulties are more prevalent mainly in the younger demographic of 25-35 age - Checks on past records must be done.
- Employment status checks for work history with the current employer must be done, along with the client's current work status. As those recruited but are on a leave (eg. Maternity Leaves) are highly likely to default.
- Cash loans are popular and are safer for the company as revolving loans have considerably high % of defaulters and can become troublesome in the long run.
- Clients who already own a realty may have problems in repaying, which might be due to existing credits. Background check on other loans must be done on such clients to prevent issues in repayment.
- While Cross-Selling is overall safe, some combinations of cross-selling have very high risk such as
  - (Cash Street(high & mid),
  - Cash x-sell) and these should be considered before approval.

RECOMMENDATIONS continued...

- Loan that were taken for **Repairs** or **Others** have high default rate, and needs to be assessed.
- Academic Status of client must be assessed and considered before approval.
- Clients who were previously refused and reapplied, checks should be made not only for their current info/status but also the reason of their previous rejection.
- Insurance and vehicle loans should have a more stringent approval process, to reduce their much higher default rate.
- While gender shouldn't be a determinant factor on who gets a loan, Men have shown tendency to have higher defaults and a stricter approval might reduce those cases.

SOURCES AND REFERENCES

- Theme and icon assets by Slides Go and Flaticon
- Seaborn - http://seaborn.pydata.org/
- Matplotlib -https://matplotlib.org/

THANK YOU!