# Credit Exploratory Data Analysis

***- Amruta Rudrasen Kapse***

# Problem Statement

- Loan providing companies find it hard to make decision about loan approval, when they receive a loan application of applicant with insufficient and non-existent credit history. And people take advantage of this scenario by being defaulter.
- Hence two types of risks are associated with the bank's decision:

1. Not approving loan to applicant who is likely to repay loan, results in loss of business.
2. Approving loan to applicant who is likely to default or not repay the loan, results in financial loss.

# Assumption & Datasets used

Assumptions:

- Columns variable with missing values more than 40% have dropped from both previous application dataset and application dataset.

Datasets Used:

- Application data : Contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
- Previous application data : contains information about the client's previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.

# Overall approach

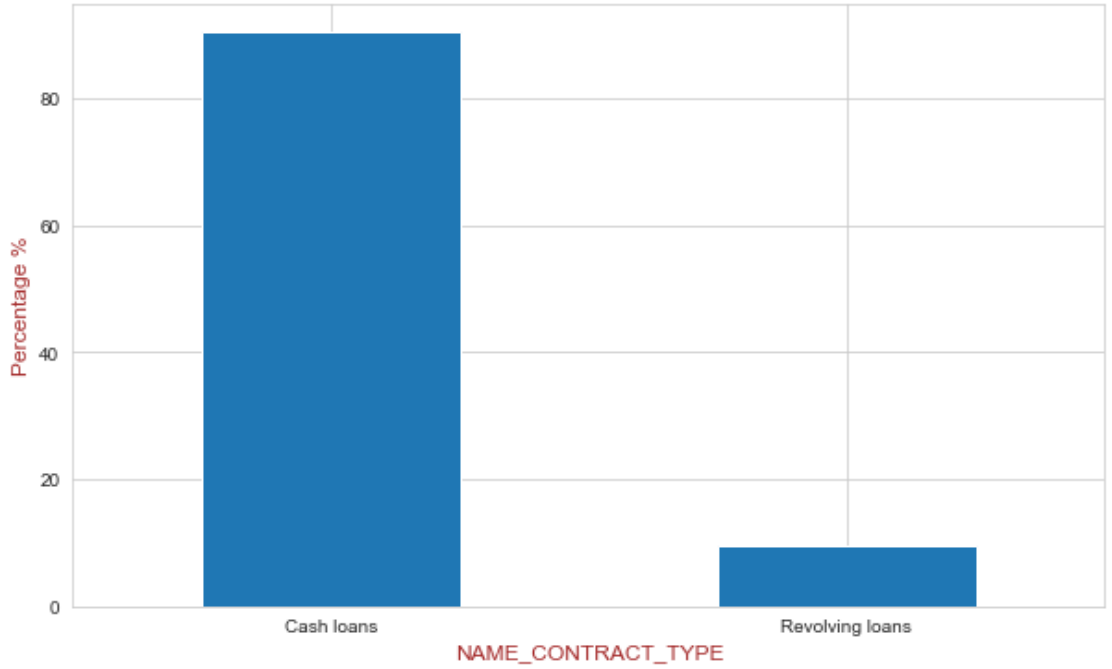
- Basic metadata check.
- Variable conversion to convenient format.
- Missing value identification[missing value identified but not treated ].
- Finding outliers in numerical variables[outliers identified but not treated ].
- Univariate and segmented univariate analysis.
- Bivariate analysis.
- Finding Top 10 correlations for the Client with payment difficulties and all other cases.

# Application Data Set

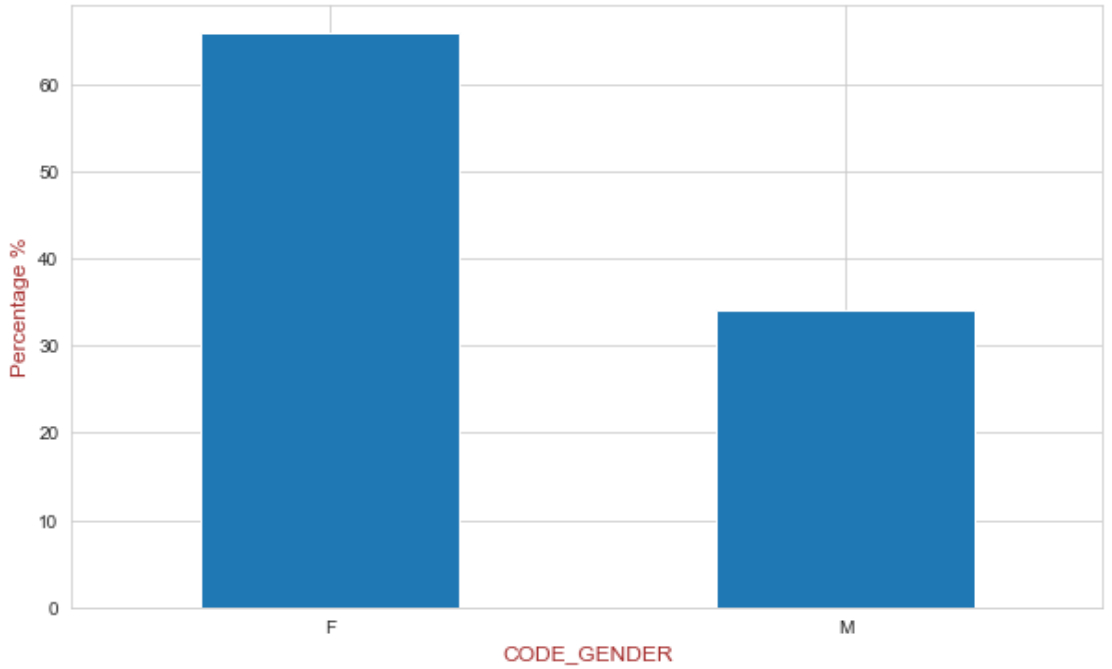
Percentage of defaulters and non-defaulters:


92%  of data is of non defaulters and 8 % of data is of client with payment difficulties.

# Univariate Analysis (1/4)

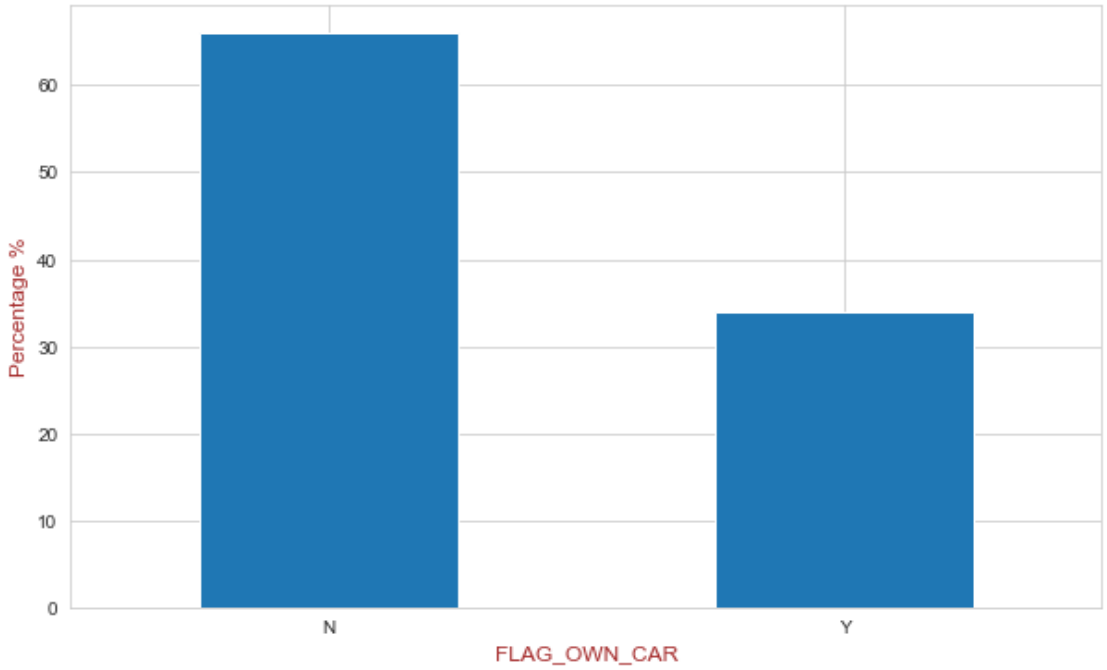



90% cash loan applications

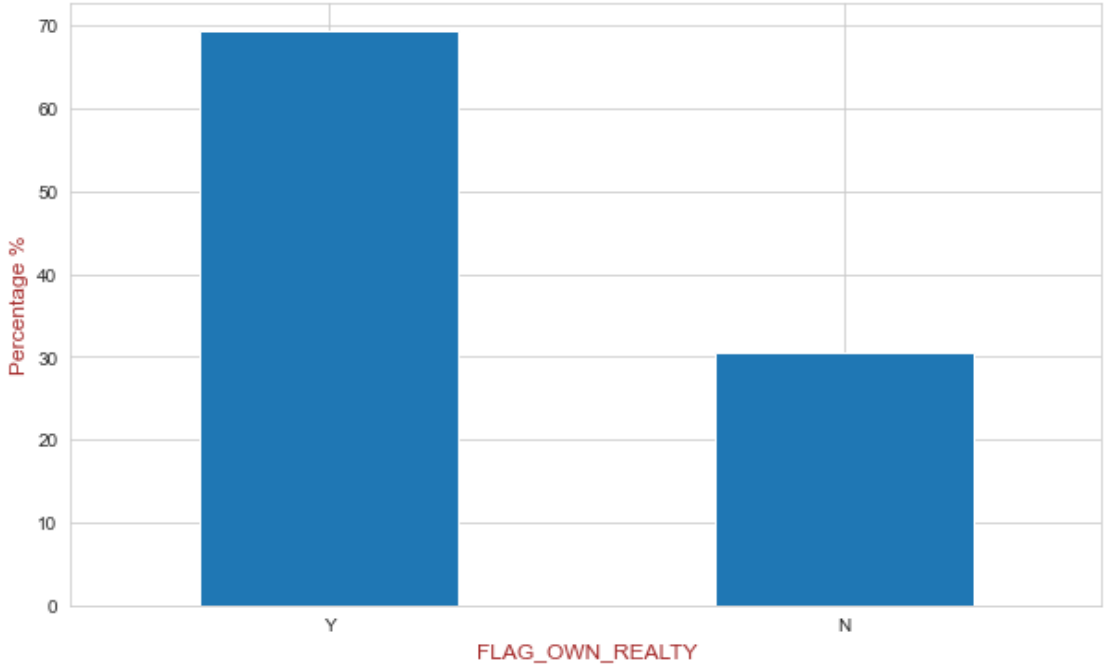



65% female applicants

# Univariate Analysis (2/4)

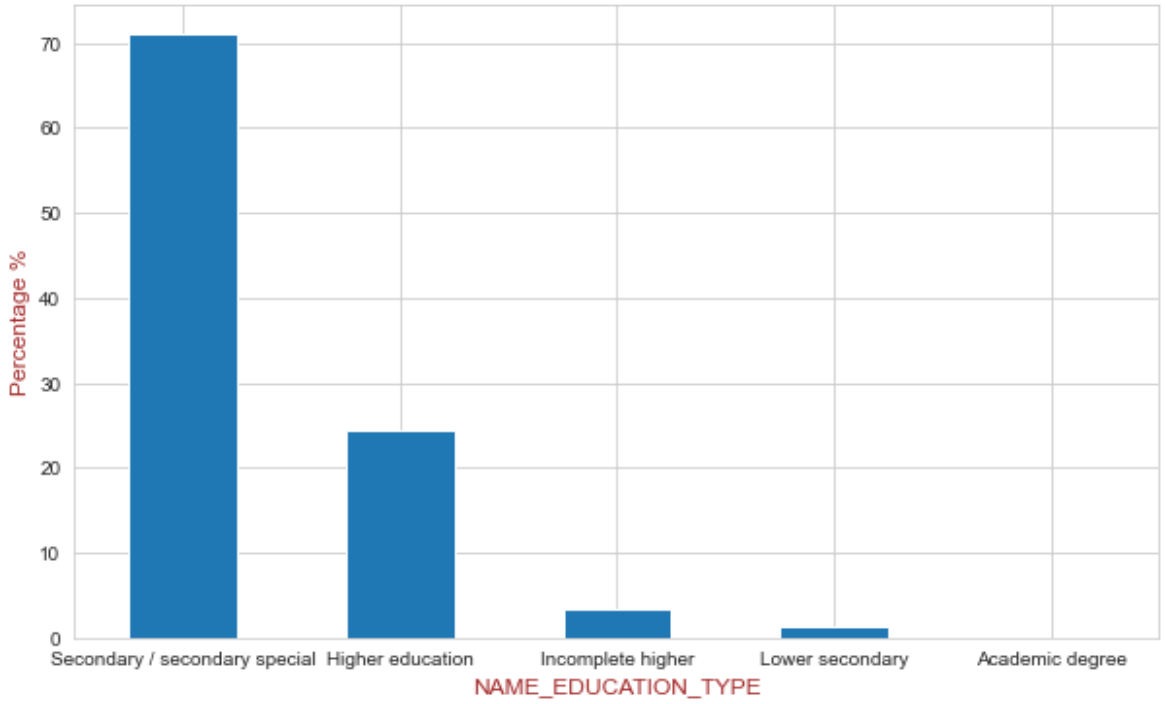



65% clients don’t own car

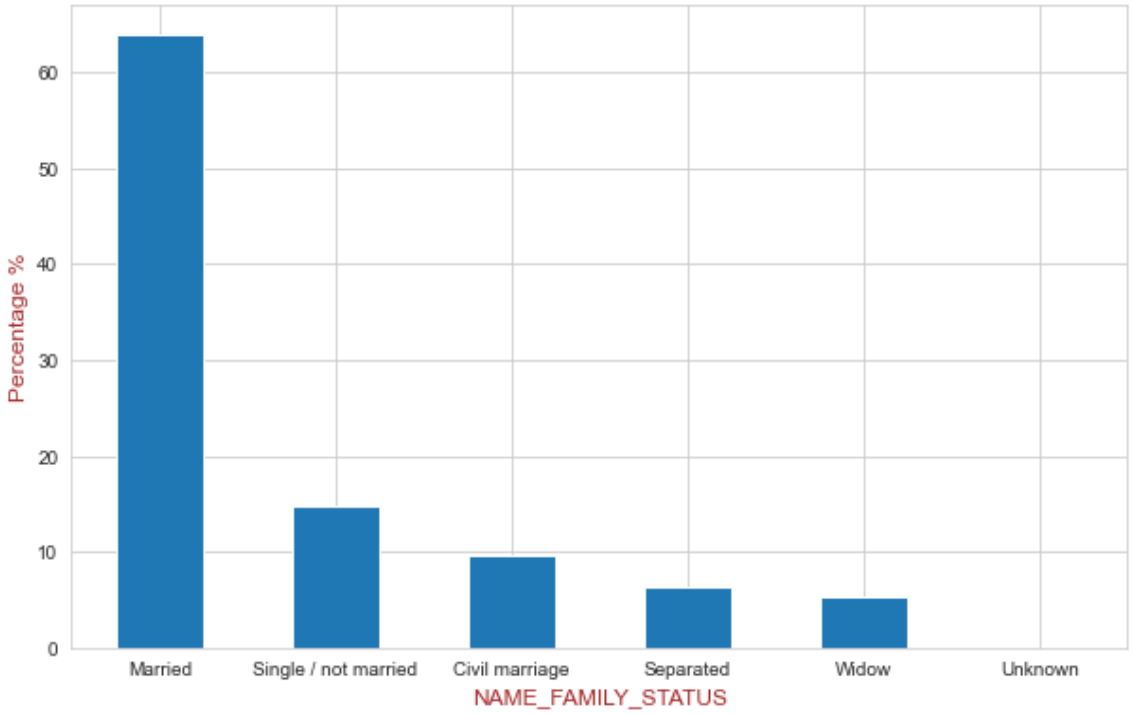



69% clients own house/apartment

# Univariate Analysis (3/4)

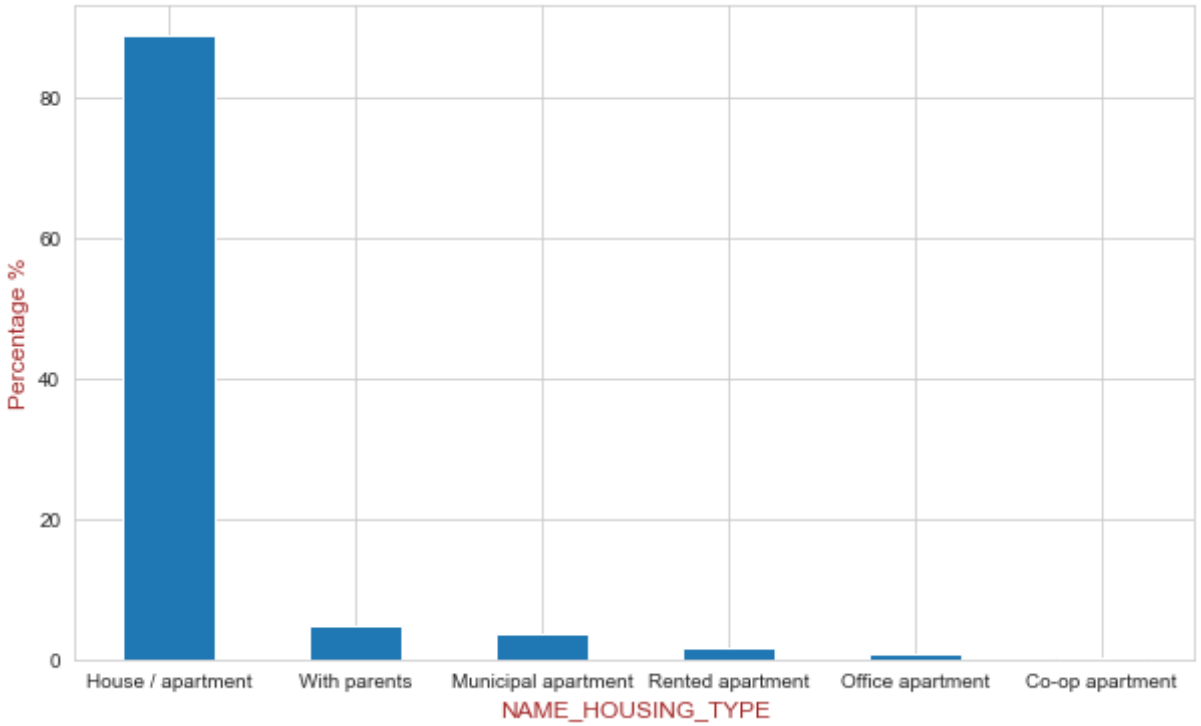



71% clients have Secondary/Secondary special education

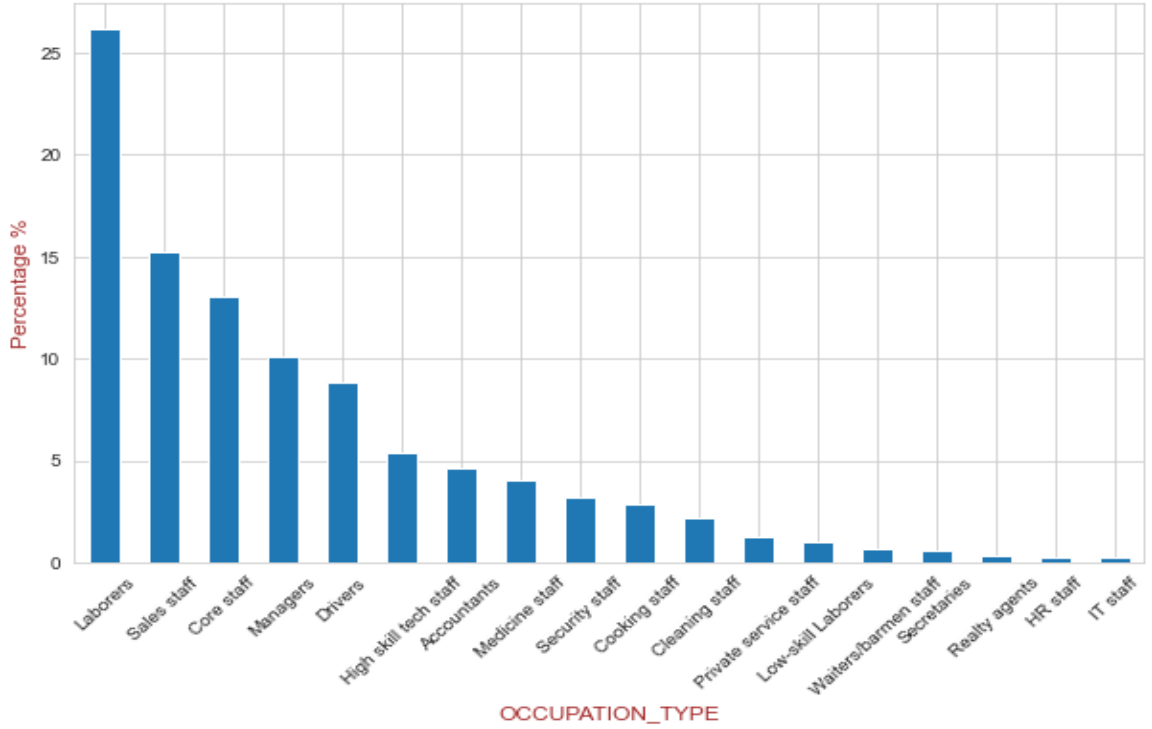



63% clients are married

# Univariate Analysis (4/4)

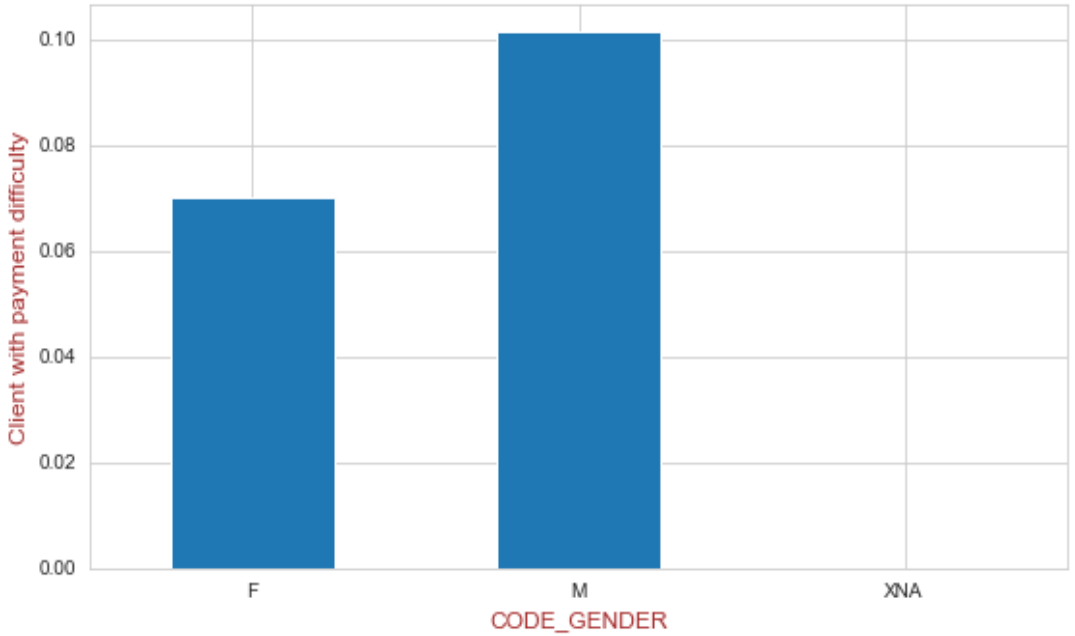



88% clients live in house/apartment



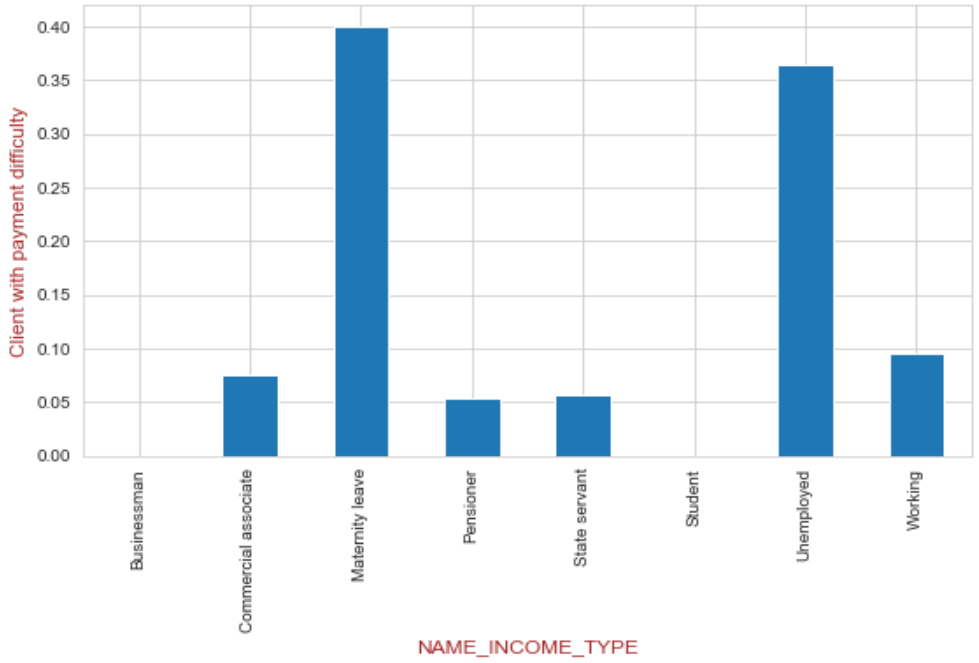

26% clients are laborers

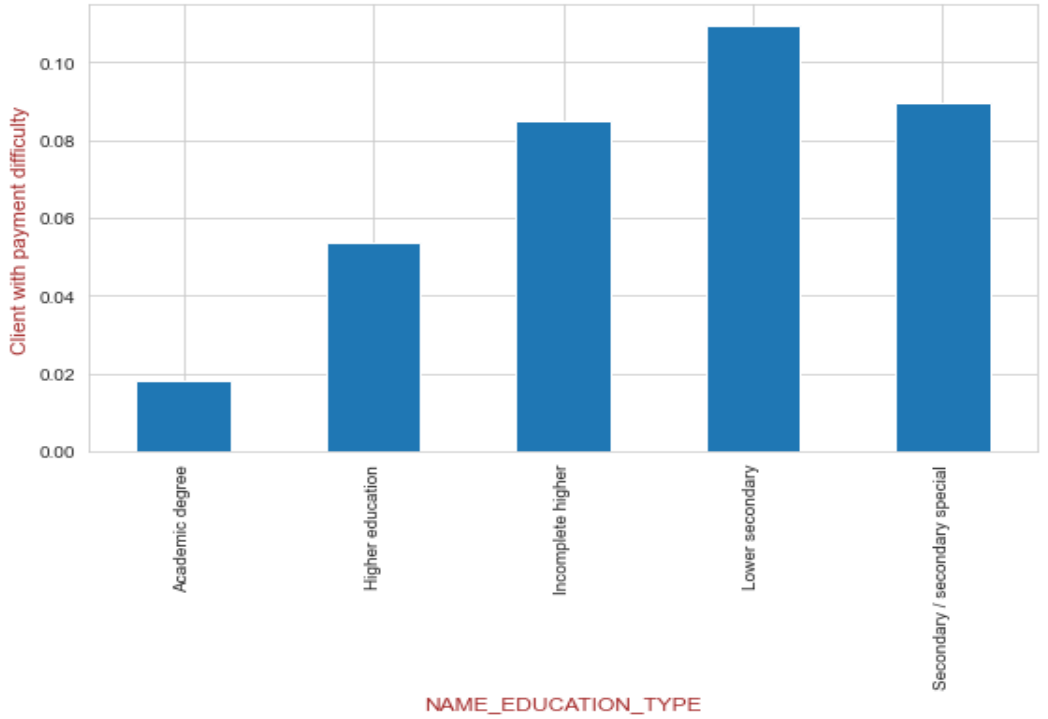
# Segmented Univariate Analysis (1/7)


- Cash loans have high default rate.
- Male clients have high default rate than female clients.

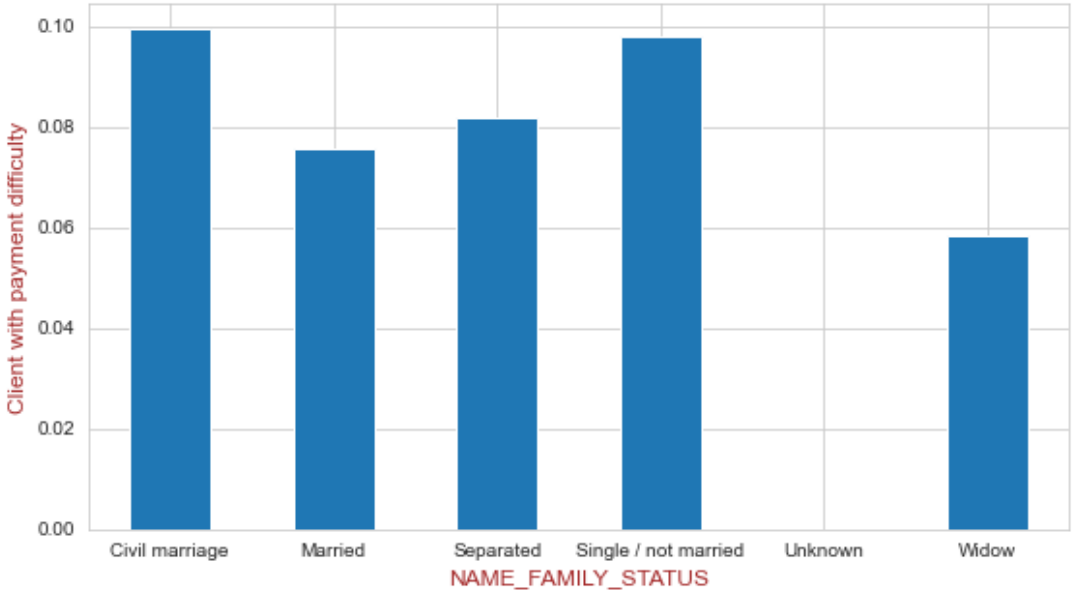
# Segmented Univariate Analysis (2/7)

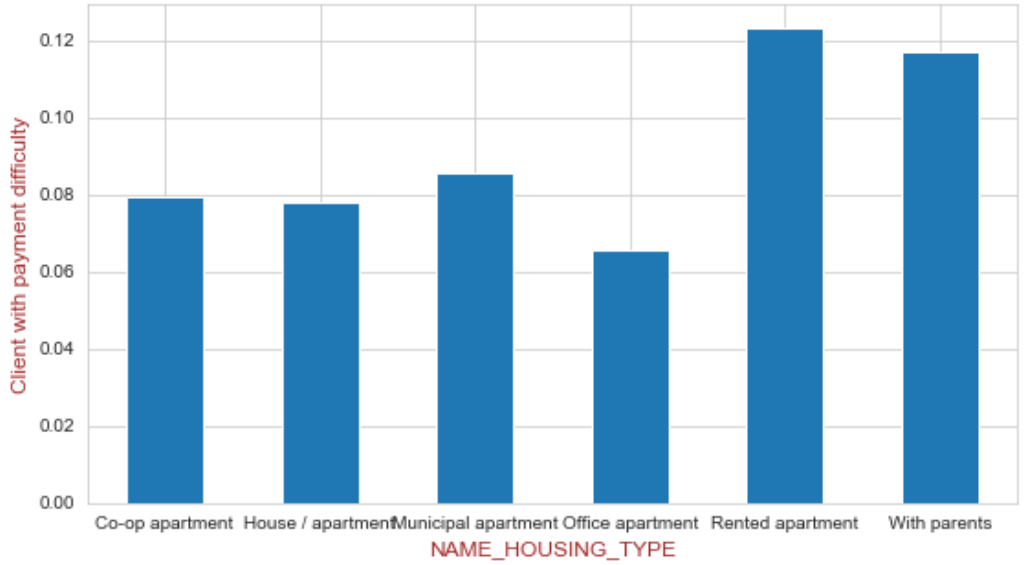



- Unemployed clients and clients on maternity leave have high default rate.
- Clients with lower secondary education have high default rate.

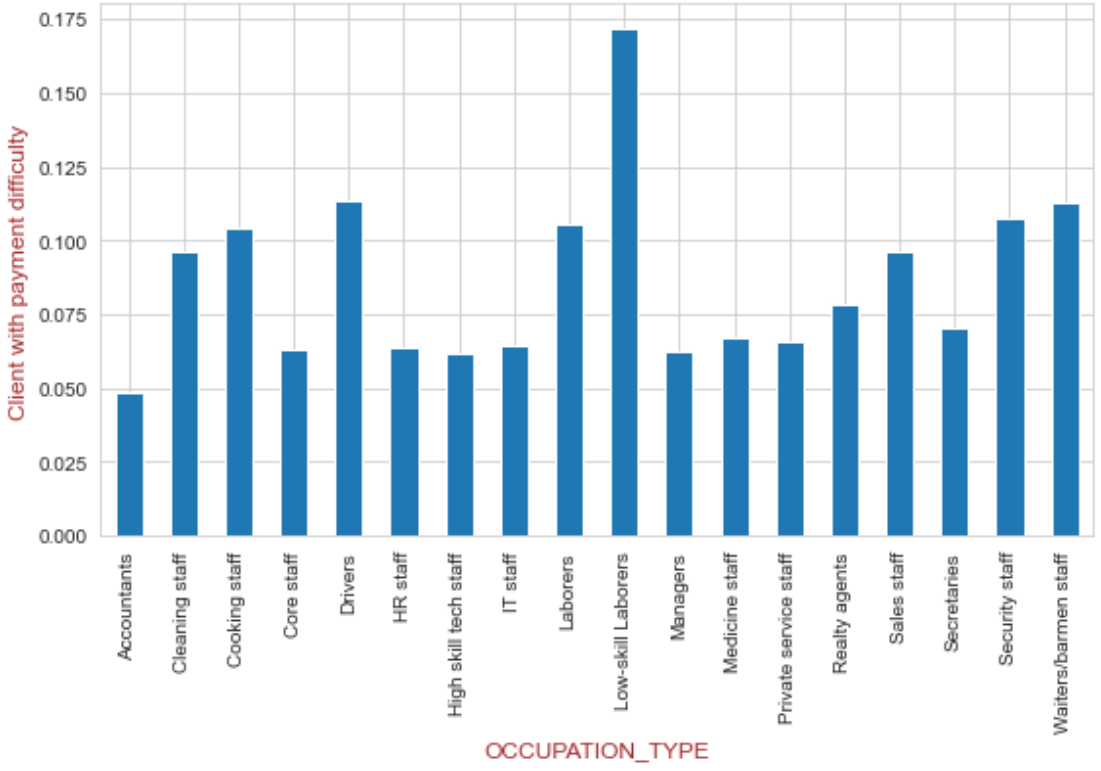
# Segmented Univariate Analysis (3/7)

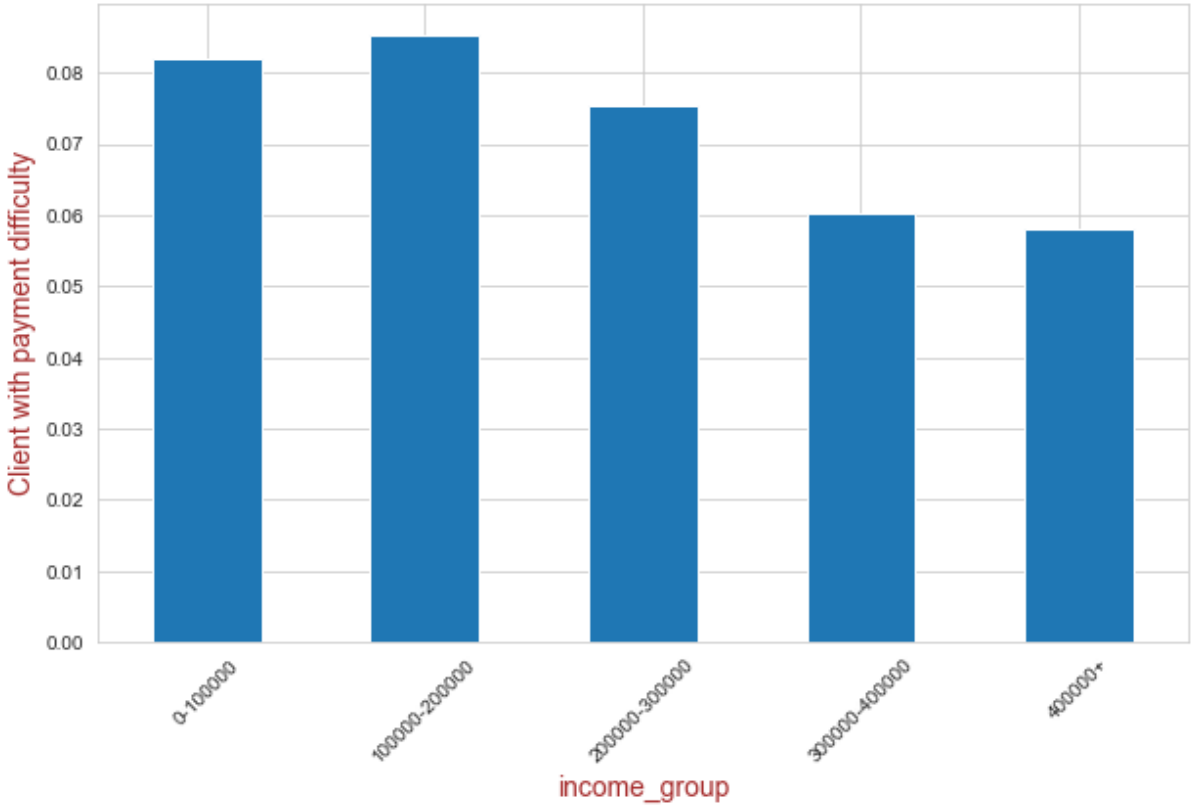



- Clients who are single and civil married have high default rate.
- Clients who live with parents and live in rented apartment have high default rate.

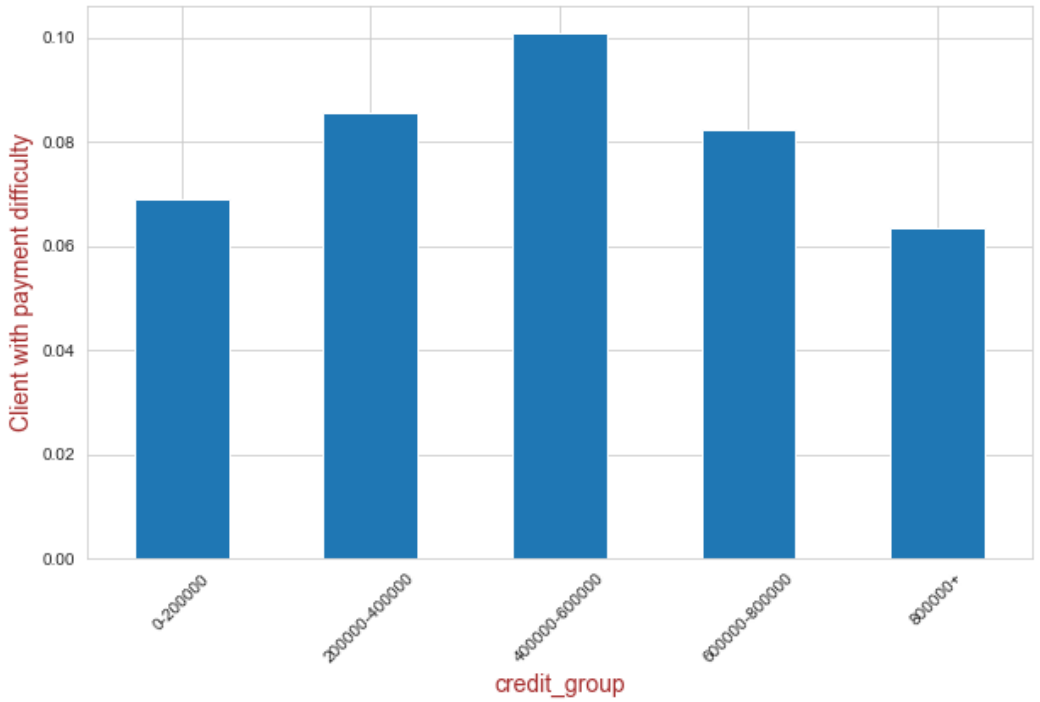
# Segmented Univariate Analysis (4/7)

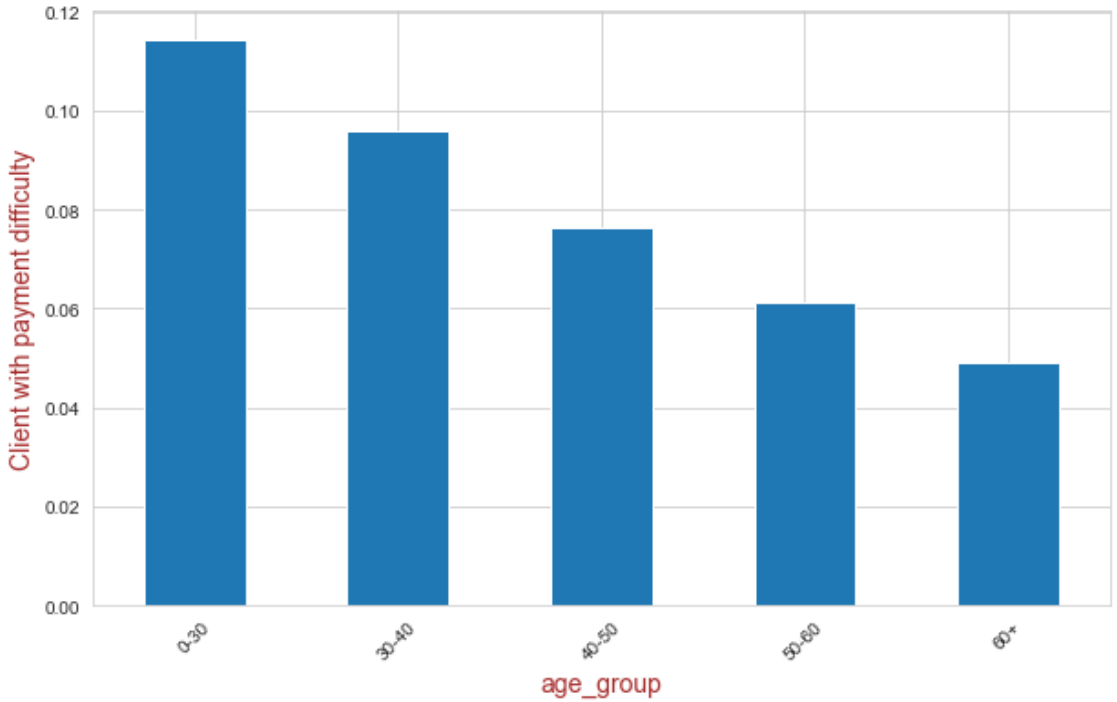



- Clients who are low skill laborers have high default rate.
- Clients who have total income upto 200000 have default rate slightly higher.

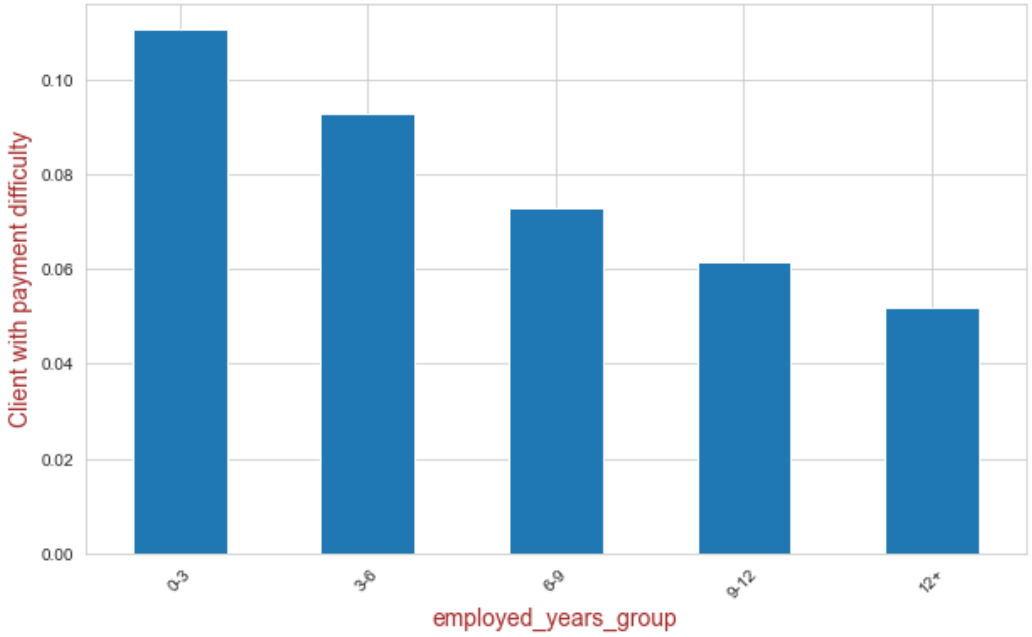
# Segmented Univariate Analysis (5/7)

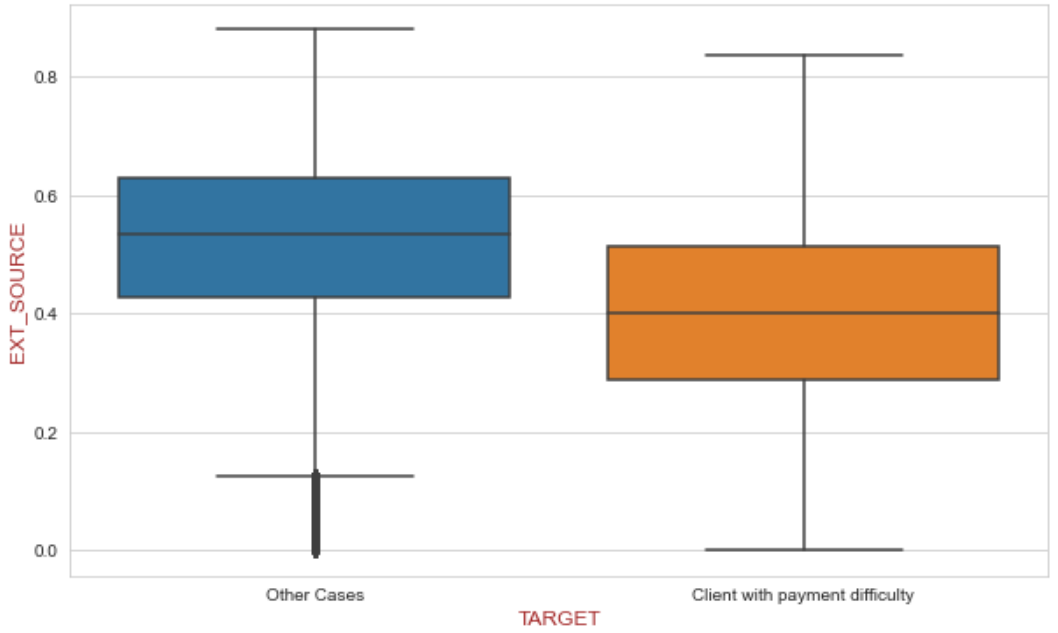



- Clients with credit amount between 400000-600000 have more default rate.
- Clients with age less than 30 defaulted more.

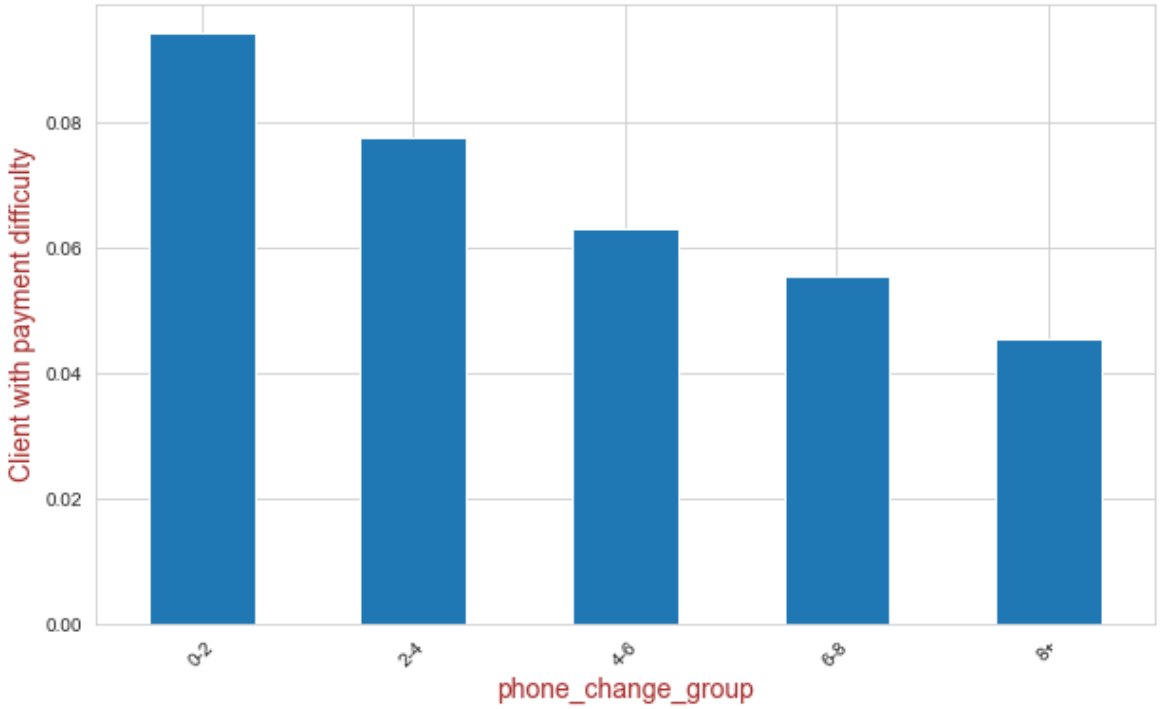
# Segmented Univariate Analysis (6/7)




- Clients with less professional experience have high default rate.
- Clients with credit score less than 0.53 have high default rate

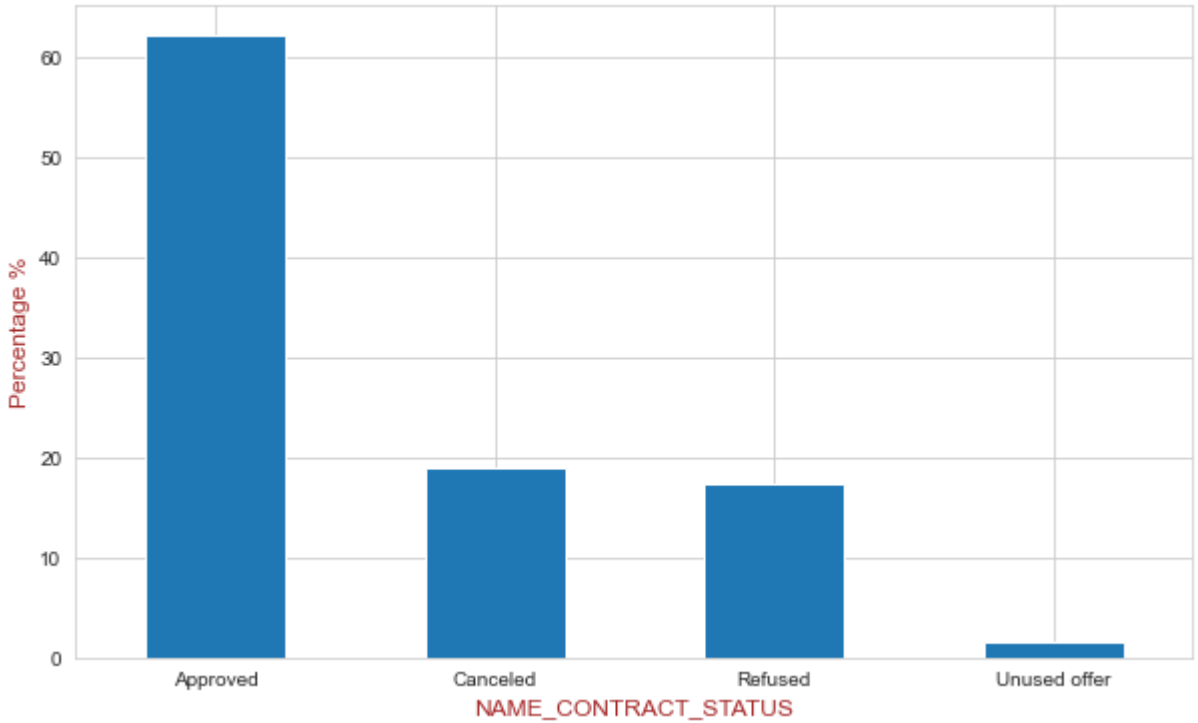
# Segmented Univariate Analysis (7/7)


- Type 3 rating of region considering city have high default rate.
- Clients who changed their phone number between 0-2 years have high default rate.

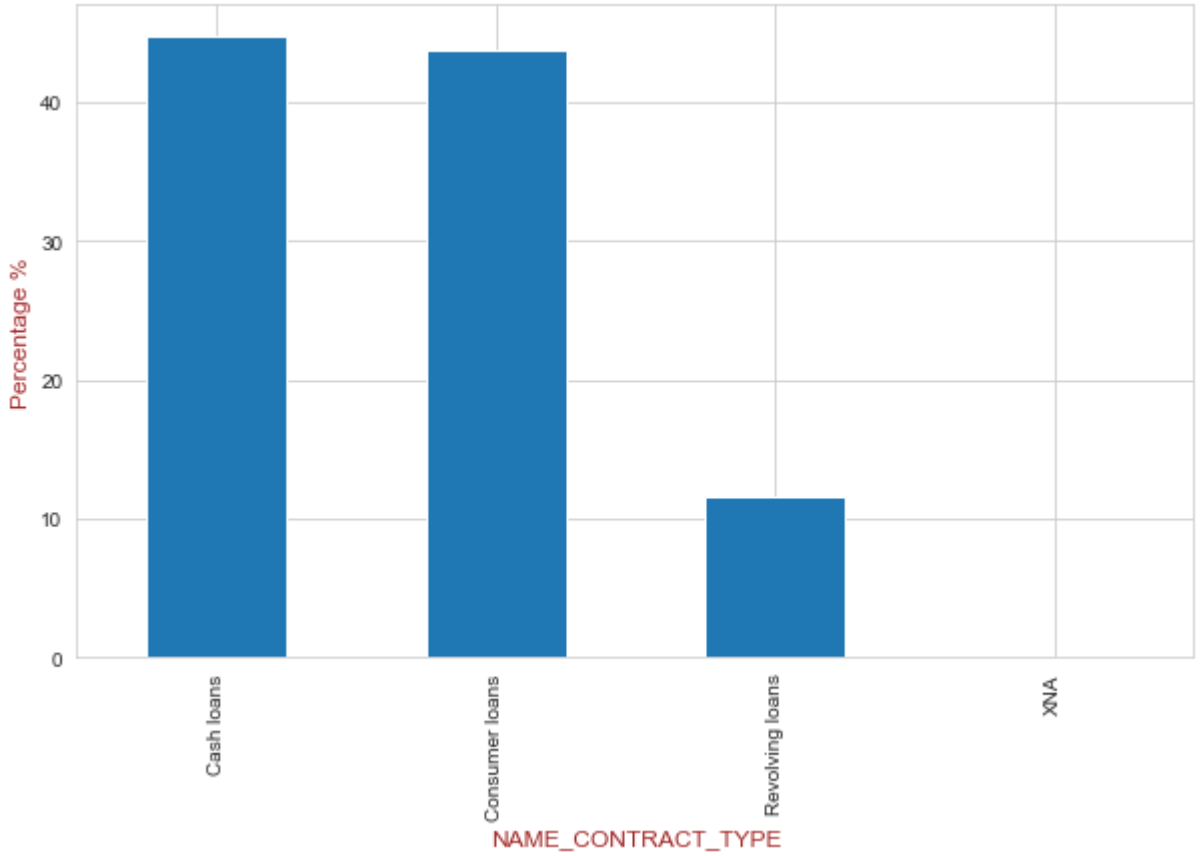
# Previous Application Data Set

Percentage of application status:

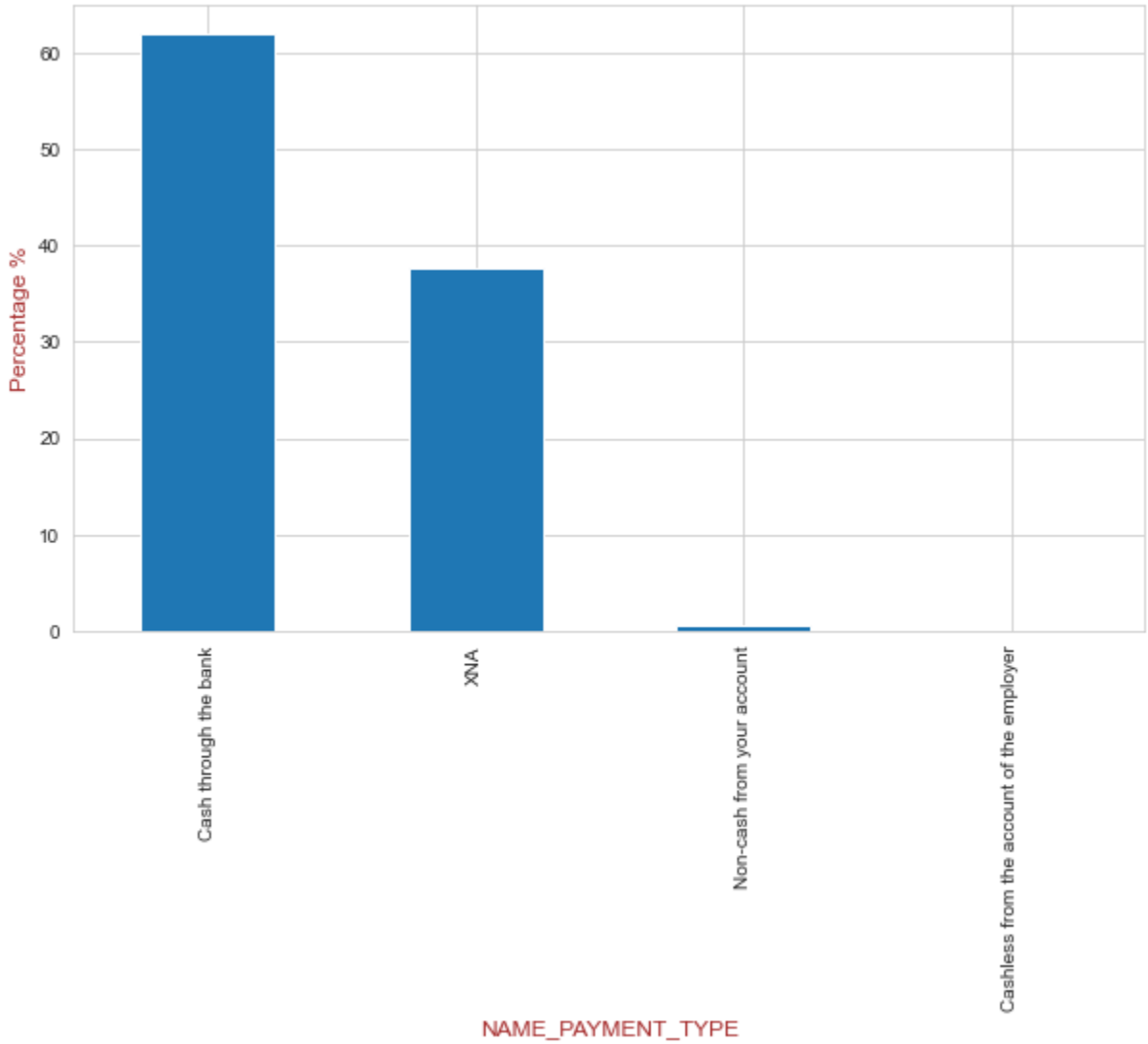



62% of data is of approved application status.

# Univariate Analysis (1/3)

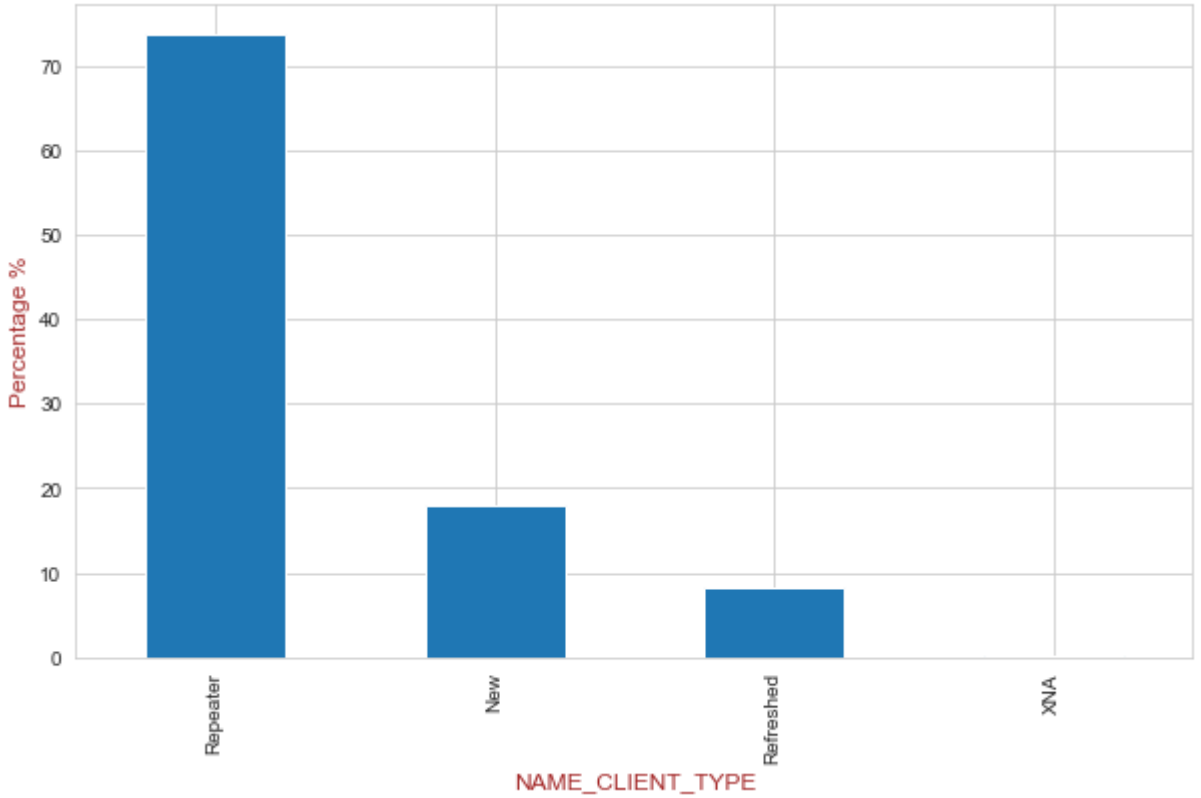



44 % cash loan applications  and 43% consumer loan applications

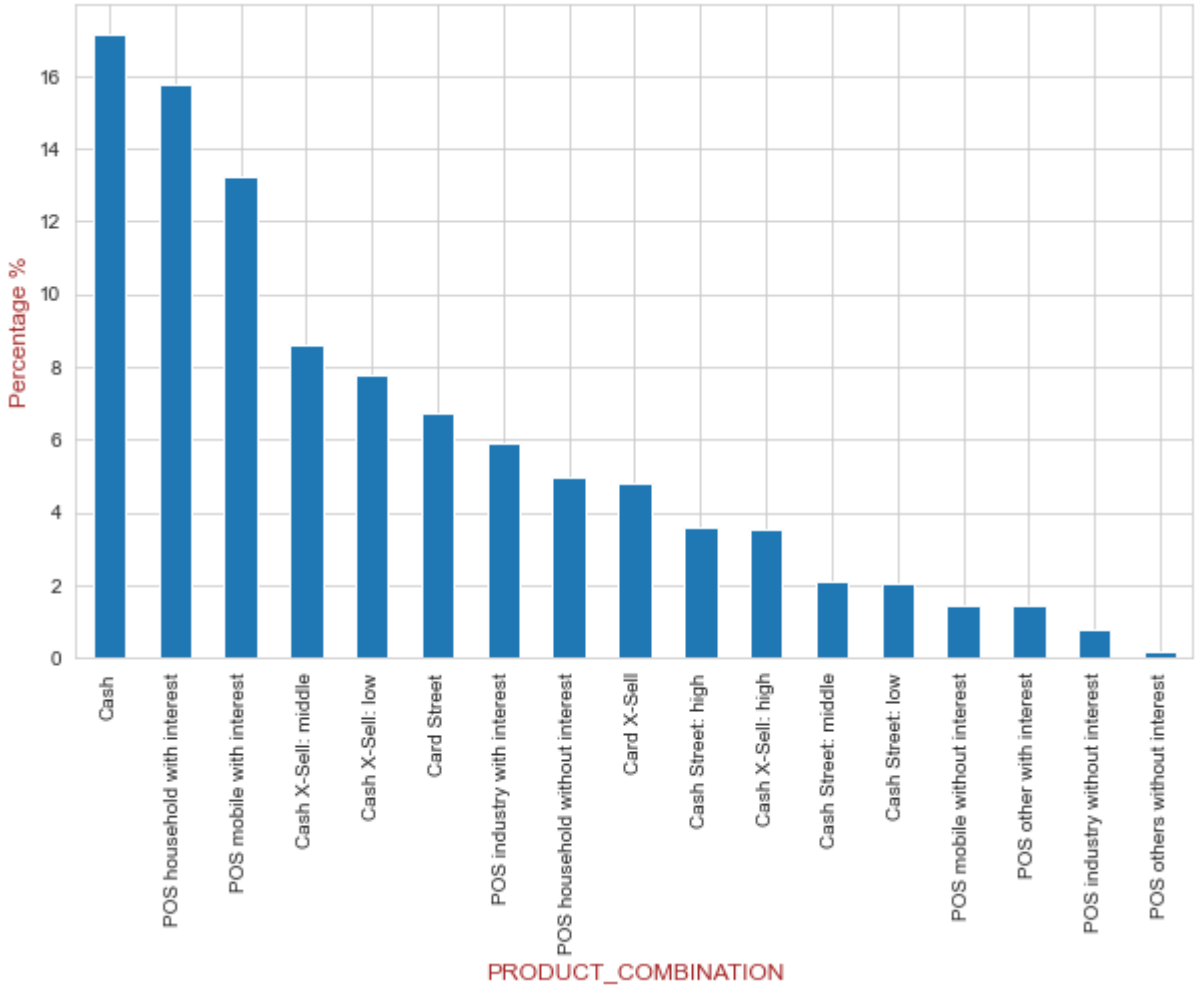



61 % payment type is cash through bank

# Univariate Analysis (2/3)

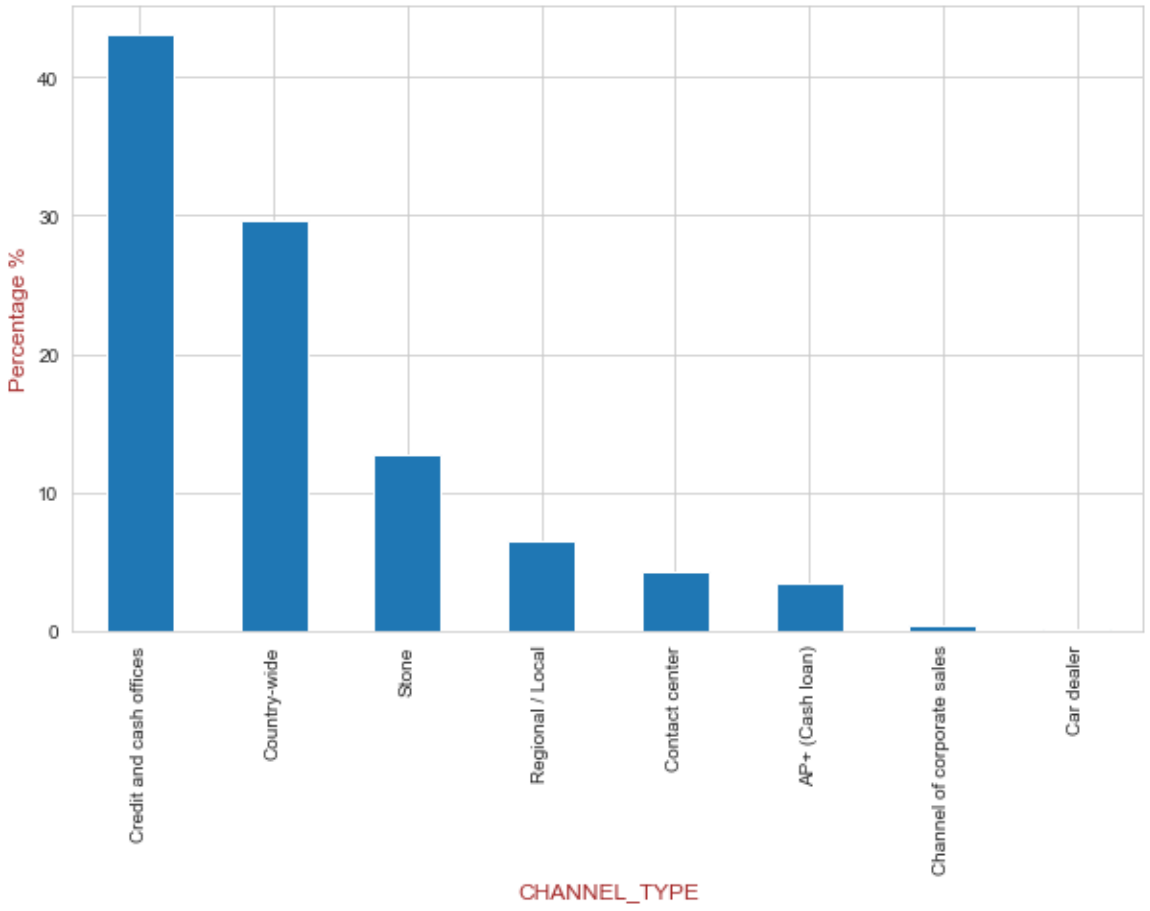



73% loan application are of repeating clients

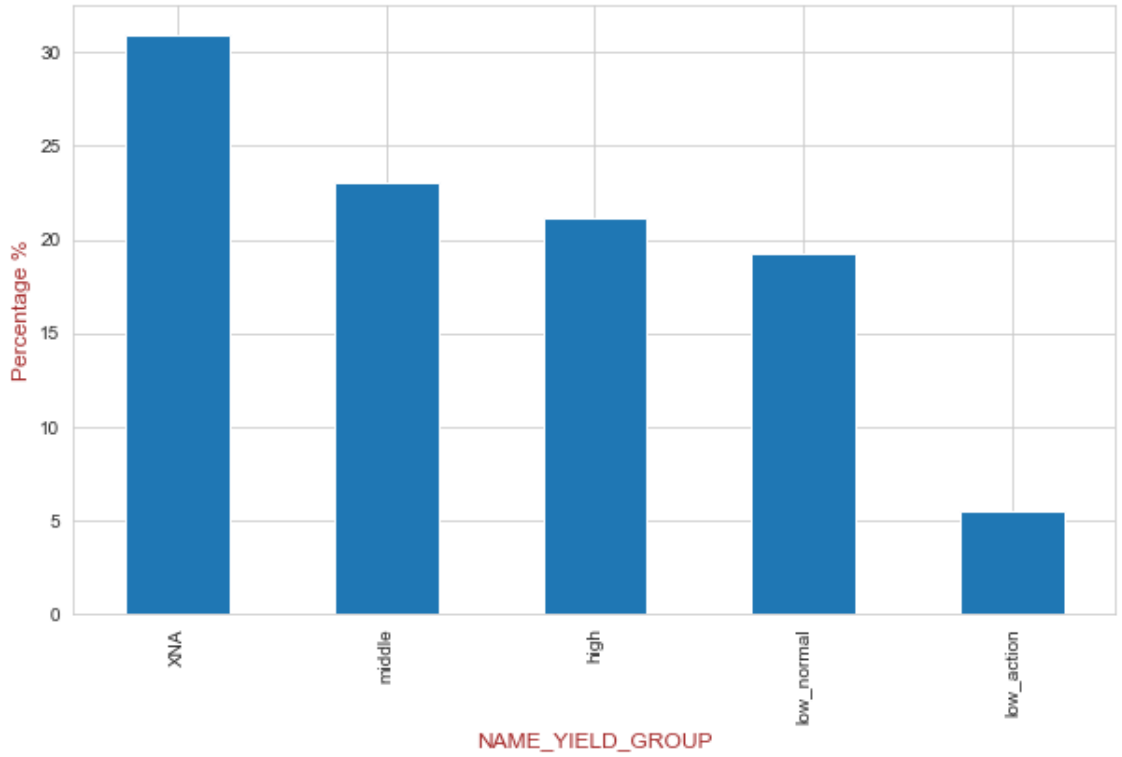



17% application is of cash, and 15% is for POS household with interest

# Univariate Analysis (3/3)

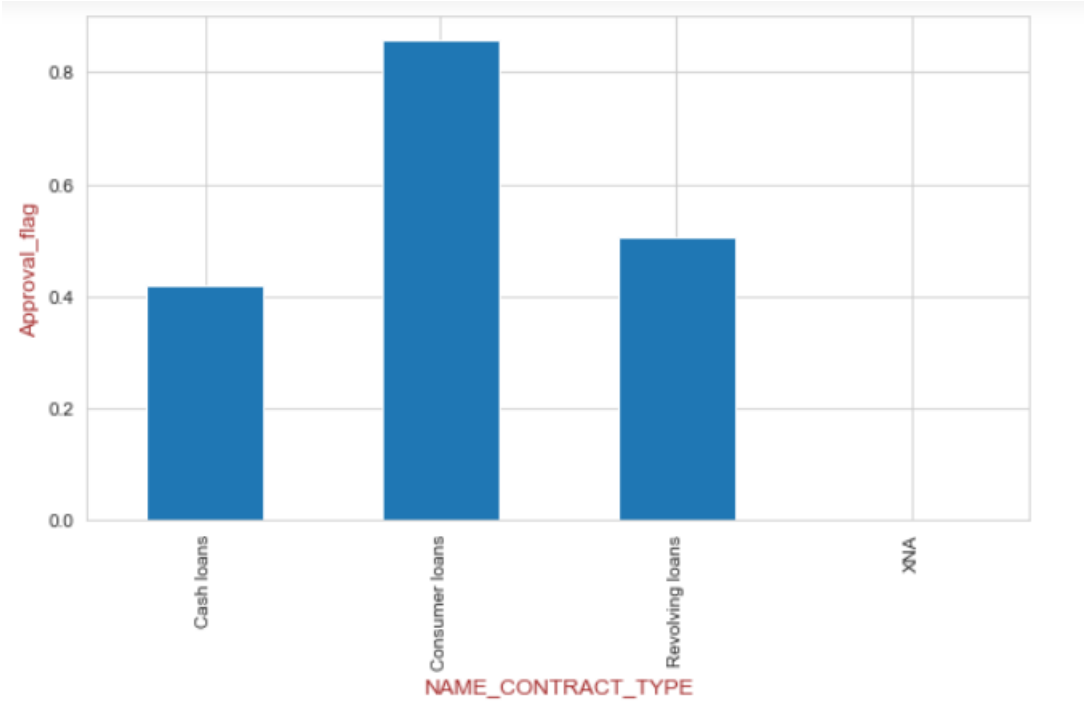


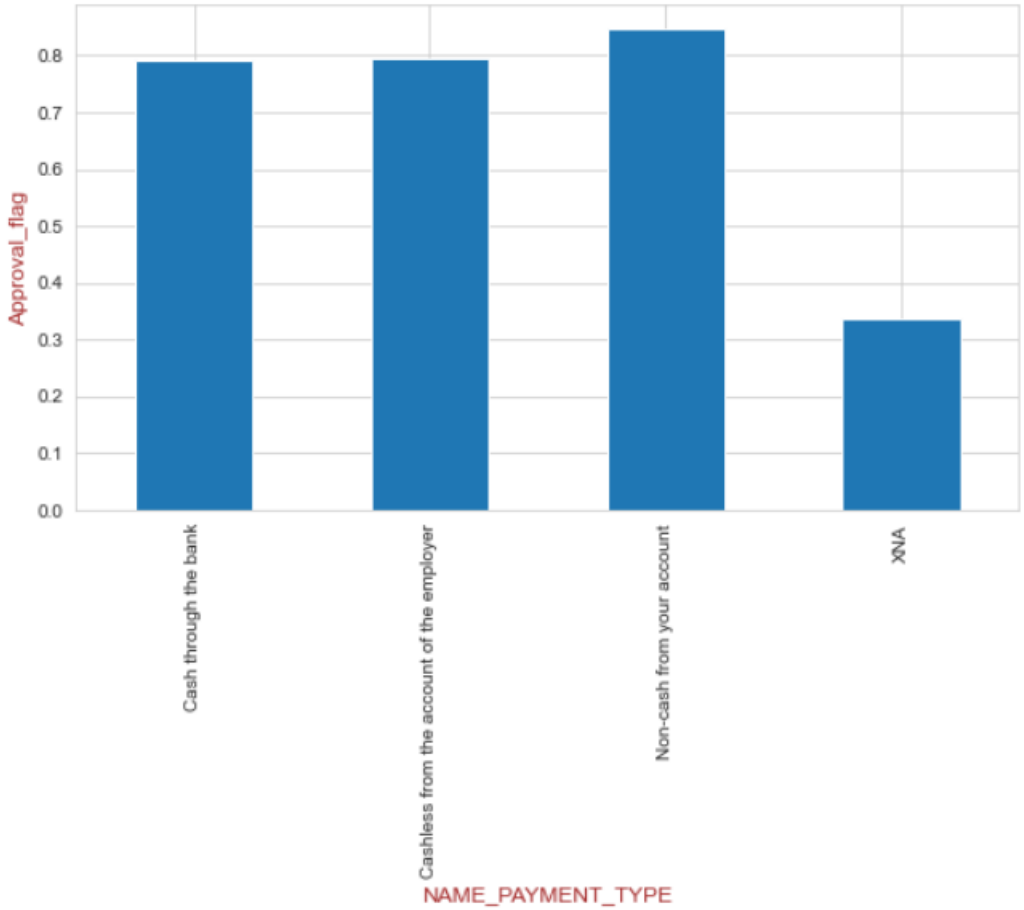

43 % loan application are through Credit and cash offices




23% application is of middle interest rate

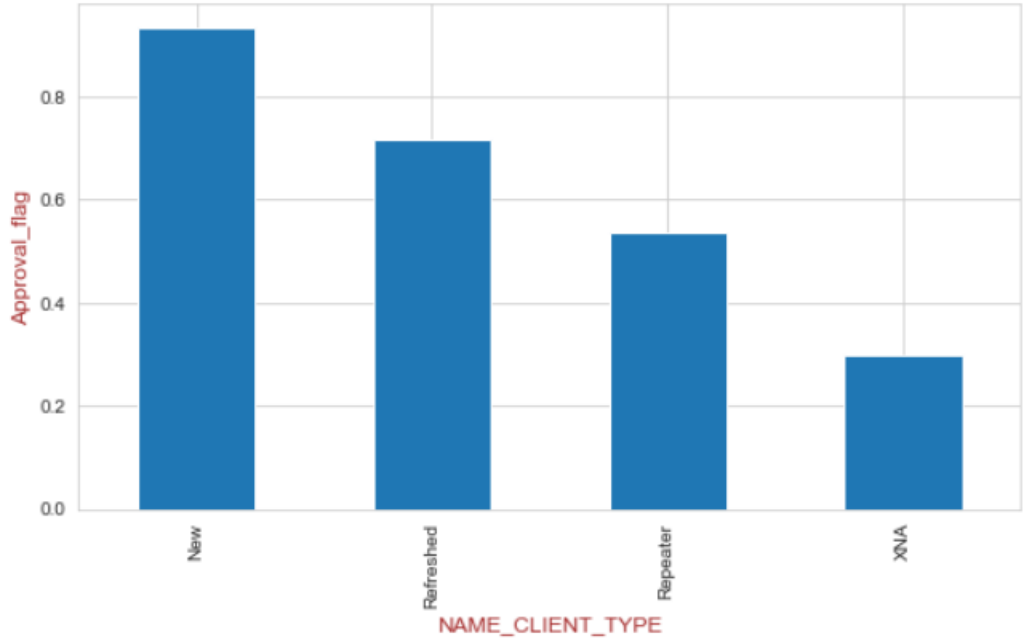
# Segmented Univariate Analysis (1/3)

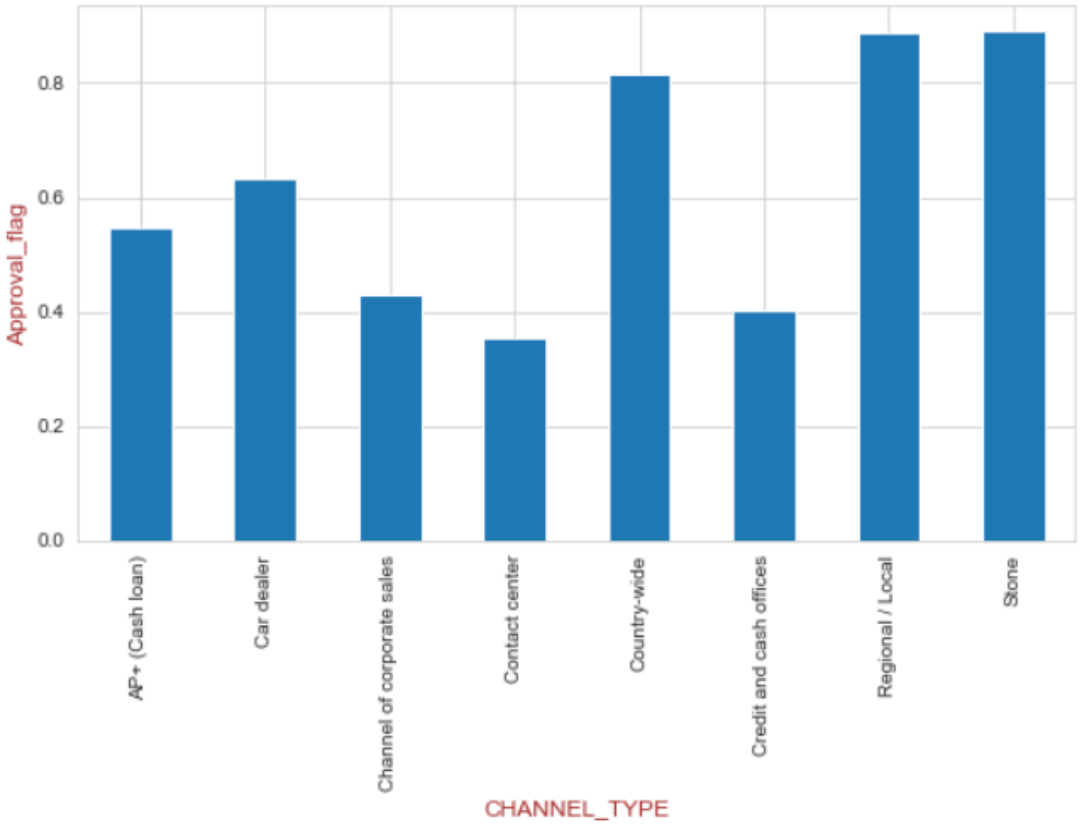



- Consumer loans have higher rate of approval.
- In approved applications cash through the bank is most used payment option.

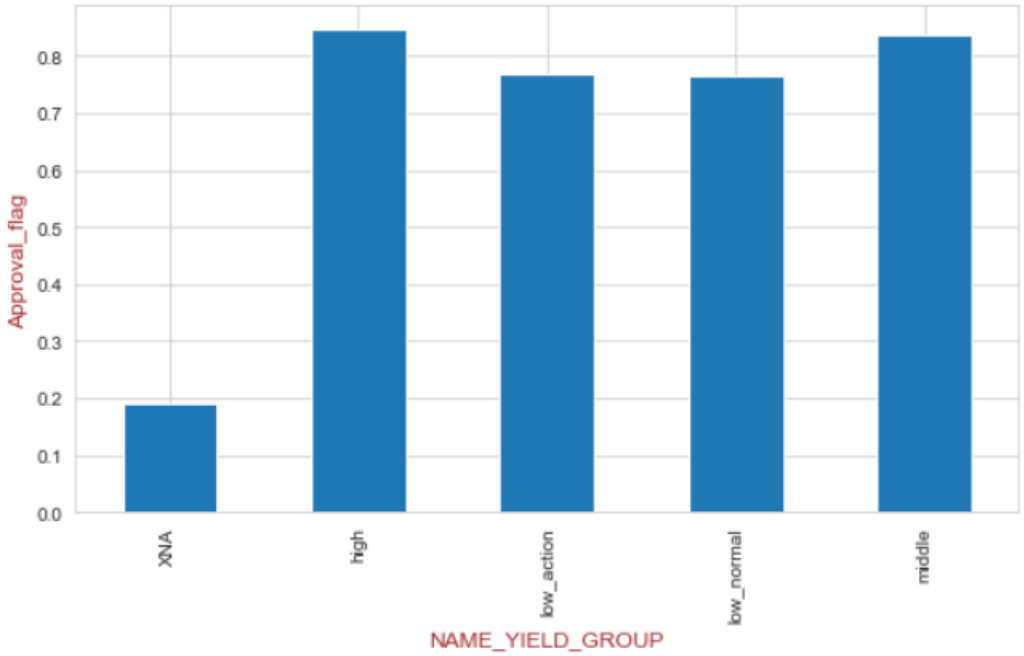
# Segmented Univariate Analysis (2/3)

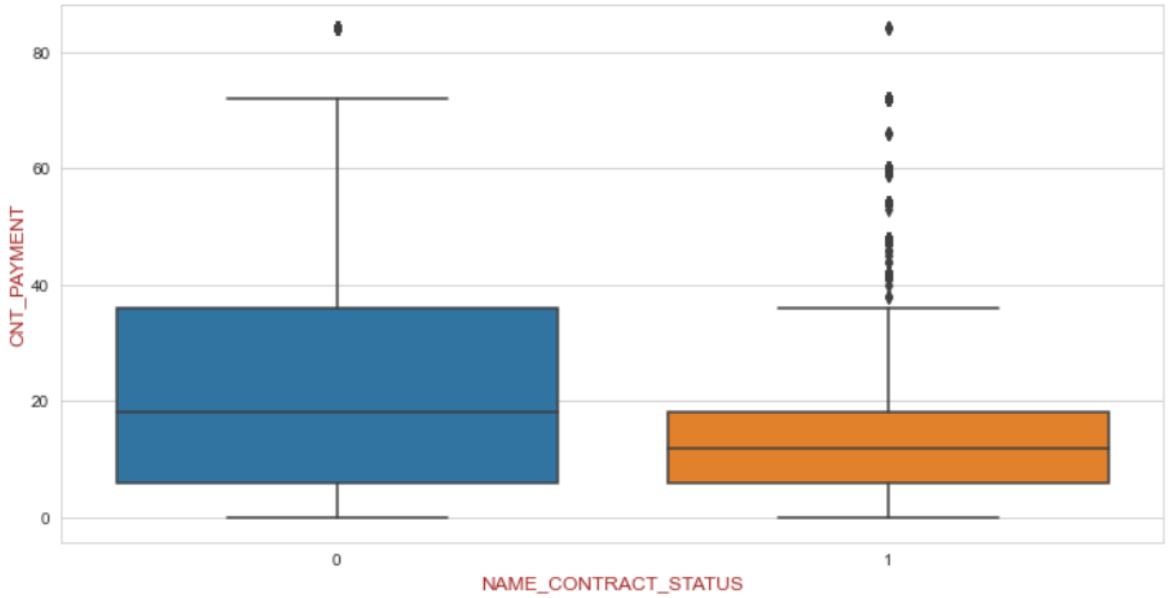



- New client type have higher rate of approval.
- Regional/local and stone channels type have higher rate of approval

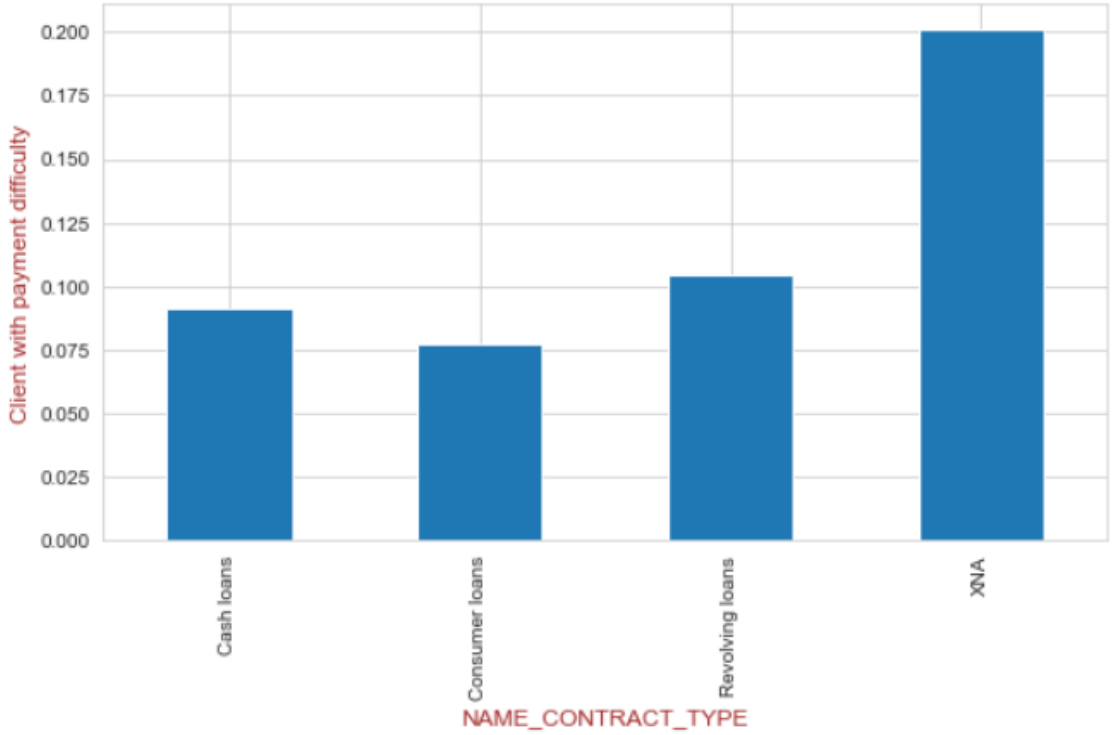
# Segmented Univariate Analysis (3/3)

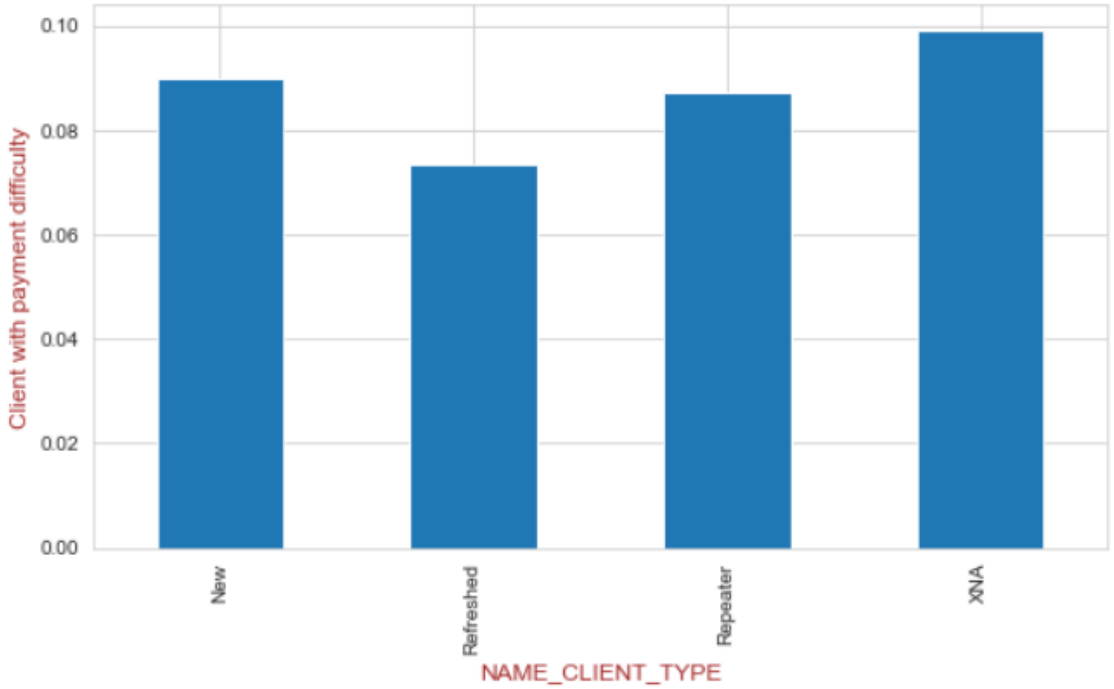



- Middle and high interest rate have more approval rate.
- Term of previous credit is widely spread in other cases and approval spread is less comparatively.

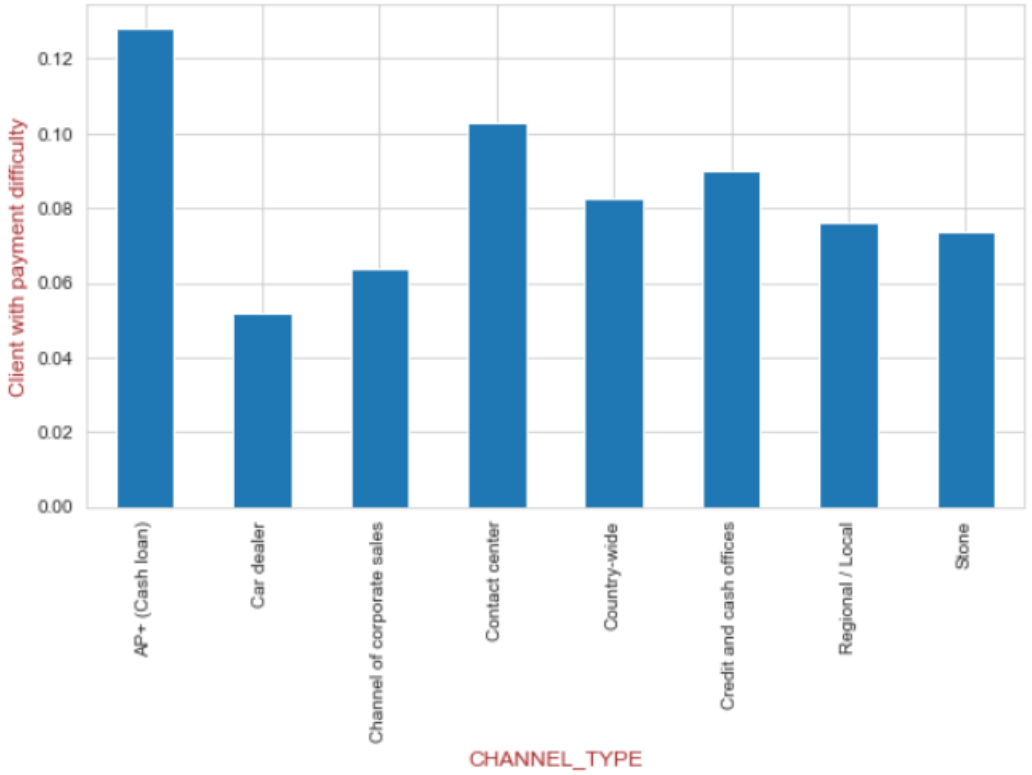
# Merged Dataset (1/3)

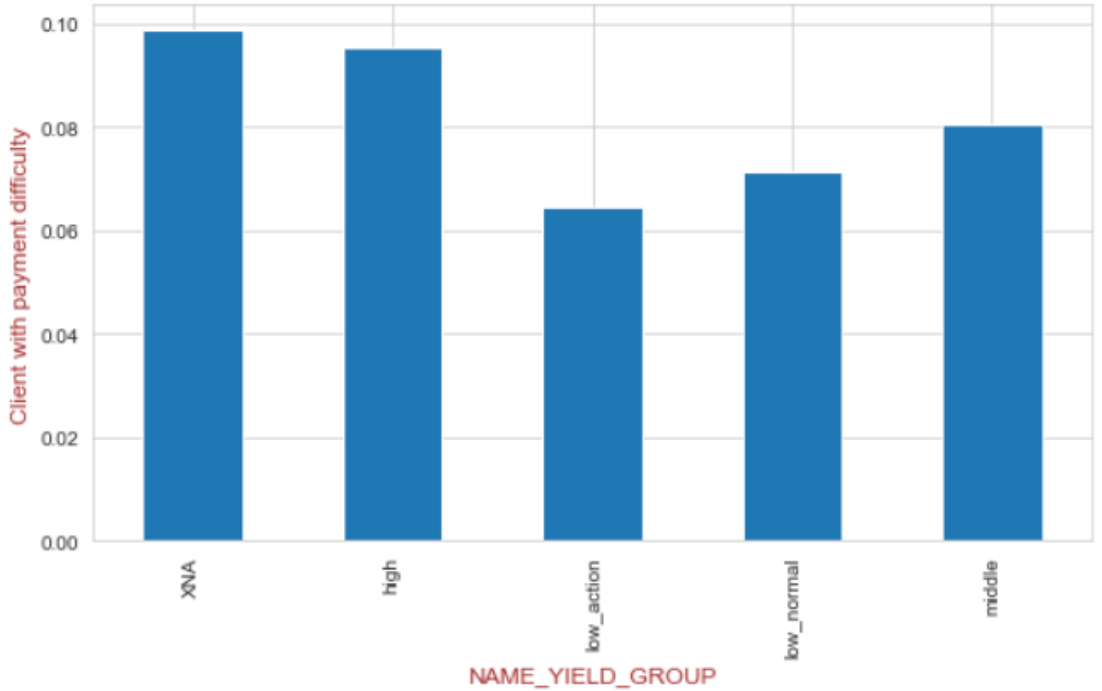



- Clients having previous loan types as revolving loan has defaulted more
- New has slightly more rate of defaulting.

# Merged Dataset (2/3)




- AP+ (Cash loan) channel application have defaulted more
- High interest rate group have defaulted more

# Merged Dataset (3/3)


- Previously refused, cancelled and unused offer have defaulted more.

# Merged Dataset :

## Top 10 correlations for target segmented data

Target = other cases:

| | | |
|---|---|---|
| OBS_60_CNT_SOCIAL_CIRCLE | OBS_30_CNT_SOCIAL_CIRCLE | 0.998578 |
| AMT_CREDIT_x | AMT_GOODS_PRICE | 0.986593 |
| AMT_CREDIT_y | AMT_APPLICATION | 0.975725 |
| REGION_RATING_CLIENT_W_CITY | REGION_RATING_CLIENT | 0.944356 |
| CNT_FAM_MEMBERS | CNT_CHILDREN | 0.878475 |
| LIVE_REGION_NOT_WORK_REGION | REG_REGION_NOT_WORK_REGION | 0.875761 |
| DEF_30_CNT_SOCIAL_CIRCLE | DEF_60_CNT_SOCIAL_CIRCLE | 0.863099 |
| LIVE_CITY_NOT_WORK_CITY | REG_CITY_NOT_WORK_CITY | 0.835641 |
| AMT_ANNUITY_y | AMT_CREDIT_y | 0.816541 |
| AMT_APPLICATION | AMT_ANNUITY_y | 0.809023 |

Target = Clients with payment difficulty:

| | | |
|---|---|---|
| OBS_30_CNT_SOCIAL_CIRCLE | OBS_60_CNT_SOCIAL_CIRCLE | 0.998378 |
| AMT_CREDIT_x | AMT_GOODS_PRICE | 0.982912 |
| AMT_APPLICATION | AMT_CREDIT_y | 0.975377 |
| REGION_RATING_CLIENT | REGION_RATING_CLIENT_W_CITY | 0.956483 |
| CNT_CHILDREN | CNT_FAM_MEMBERS | 0.886300 |
| REG_REGION_NOT_WORK_REGION | LIVE_REGION_NOT_WORK_REGION | 0.873130 |
| DEF_30_CNT_SOCIAL_CIRCLE | DEF_60_CNT_SOCIAL_CIRCLE | 0.858281 |
| AMT_CREDIT_y | AMT_ANNUITY_y | 0.840461 |
| AMT_ANNUITY_y | AMT_APPLICATION | 0.824962 |
| LIVE_CITY_NOT_WORK_CITY | REG_CITY_NOT_WORK_CITY | 0.792897 |

- Top 10 correlation for both cases is almost same

# Conclusion

- The data is imbalanced 92% of data is of non defaulter clients
- Cash loans have more default rate
- Female clients application are high in number but rate of default is high in male clients than female clients
- Businessmen default less, mor loans can be given to these people
- People with lower secondary education defaults more, clients with higher education seems safer choice for loan
- Single people and people who lives with  parents or in rented apartments default more. And less risk associated with married people
- People with low income as in case of low skilled laborers default more. Higher income less risk of defaulting
- Young people who have less employment experience defaults more, older people with much professional experience should be preferred
- People living in region of rating 3 defaults more and people who changed their number before applying for loan defaults more
- People with high interest amount have defaulted more
- Loans which are previously refused, canceled and loan offer unused have high default rate

Thank you!!