Credit Case Study

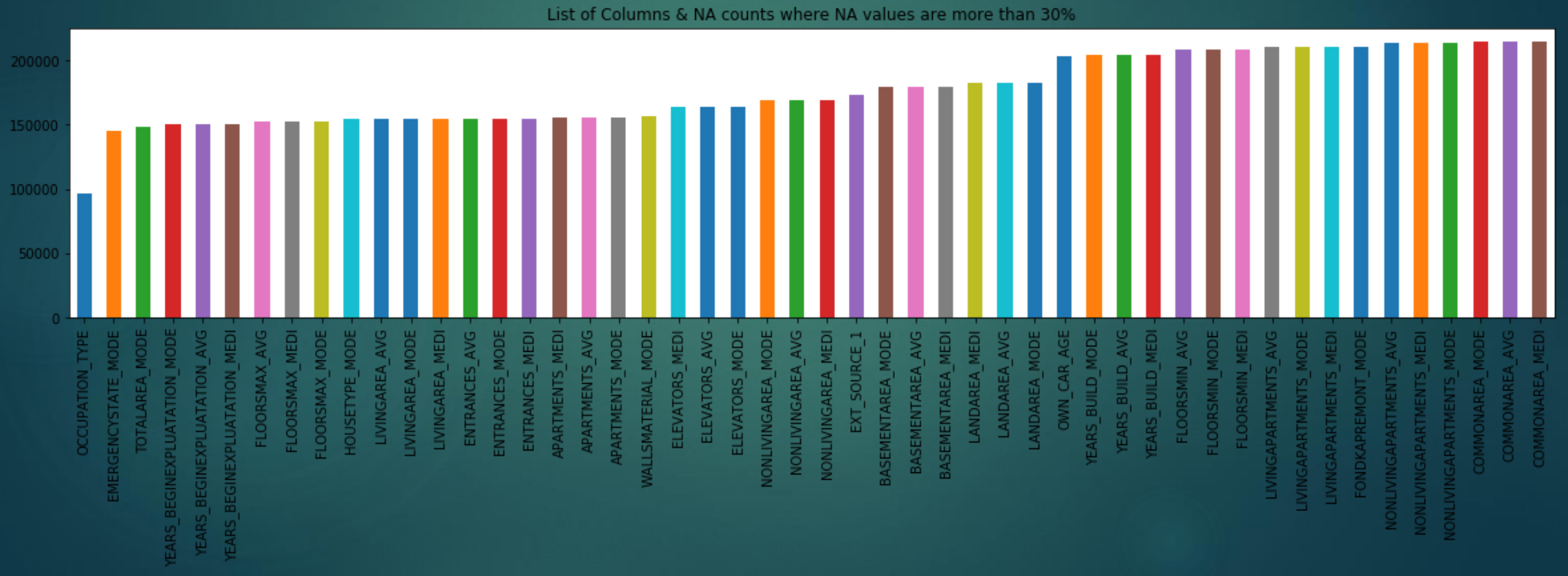
SUBMITTED BY,
KAVITHA MAHESH
JEYA BALAJI

Problem Statement-1

- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

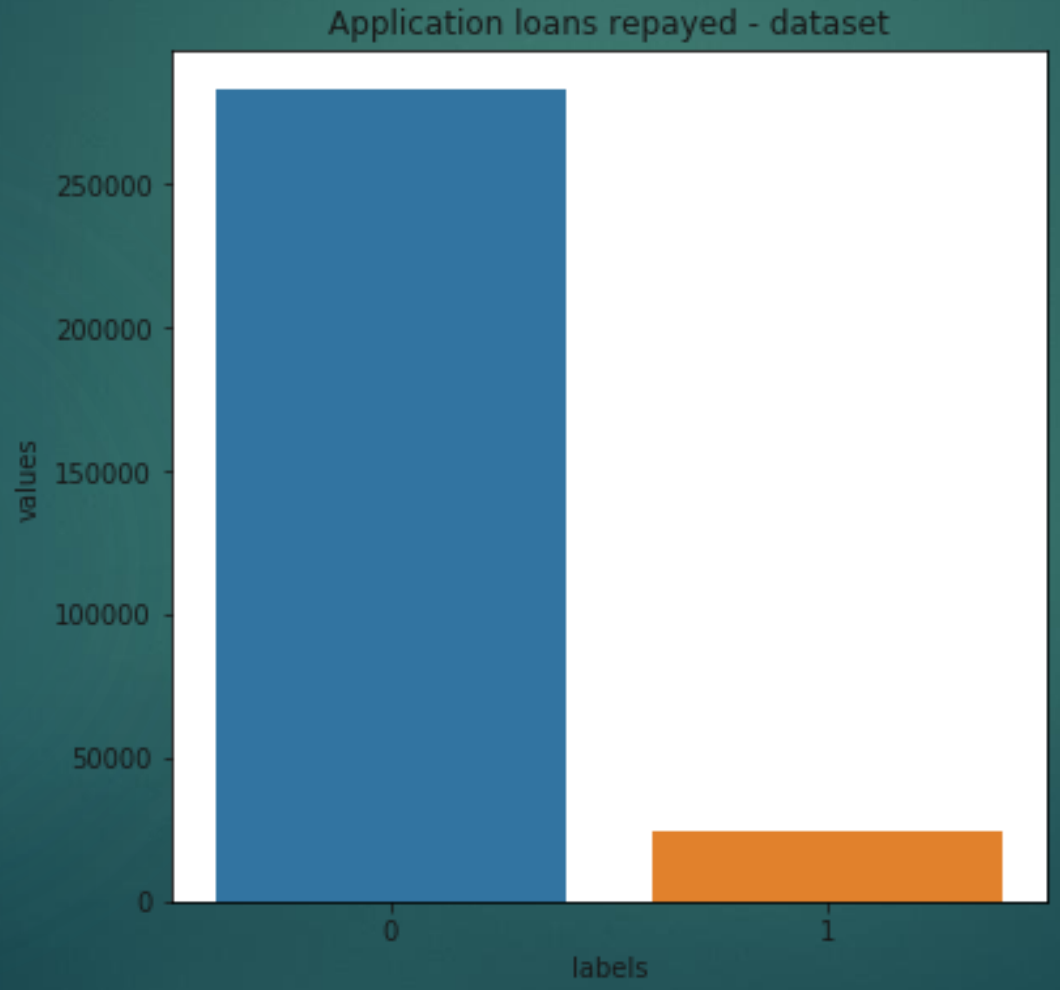
Data Exploration - Application Dataset

- Checking for data columns with less than 30% missing data
- Columns or variables with more than 30% missing data can be removed, since these variables would not account to analysis



Checking for Data unbalance

- From the below plot, it is clear that the loan defaulters are pretty less when compared to those who repay the loan, with a ratio **10:1**



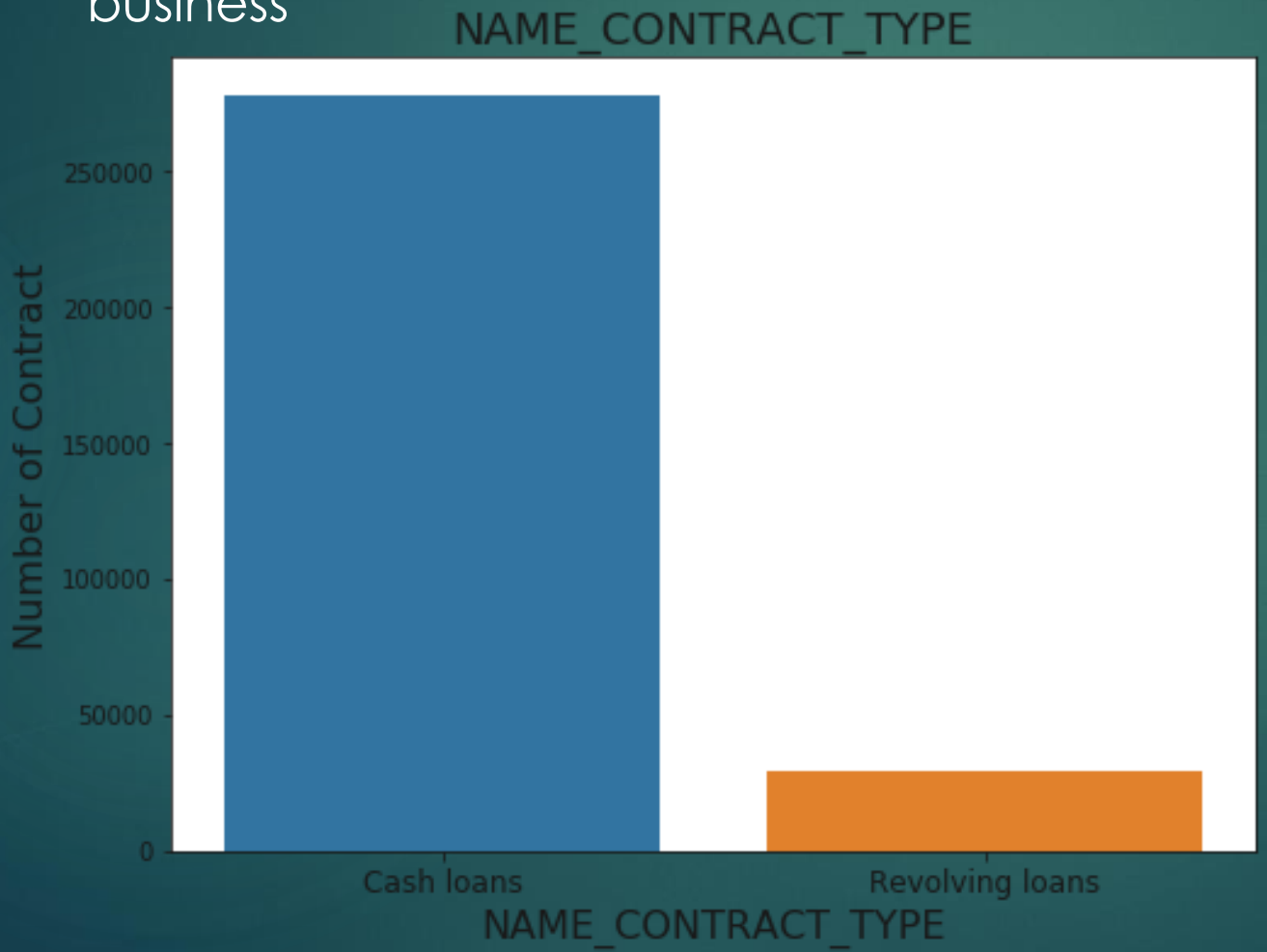
Problem Statement-2

- 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.
- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

Data Set - Analysis Application Dataset

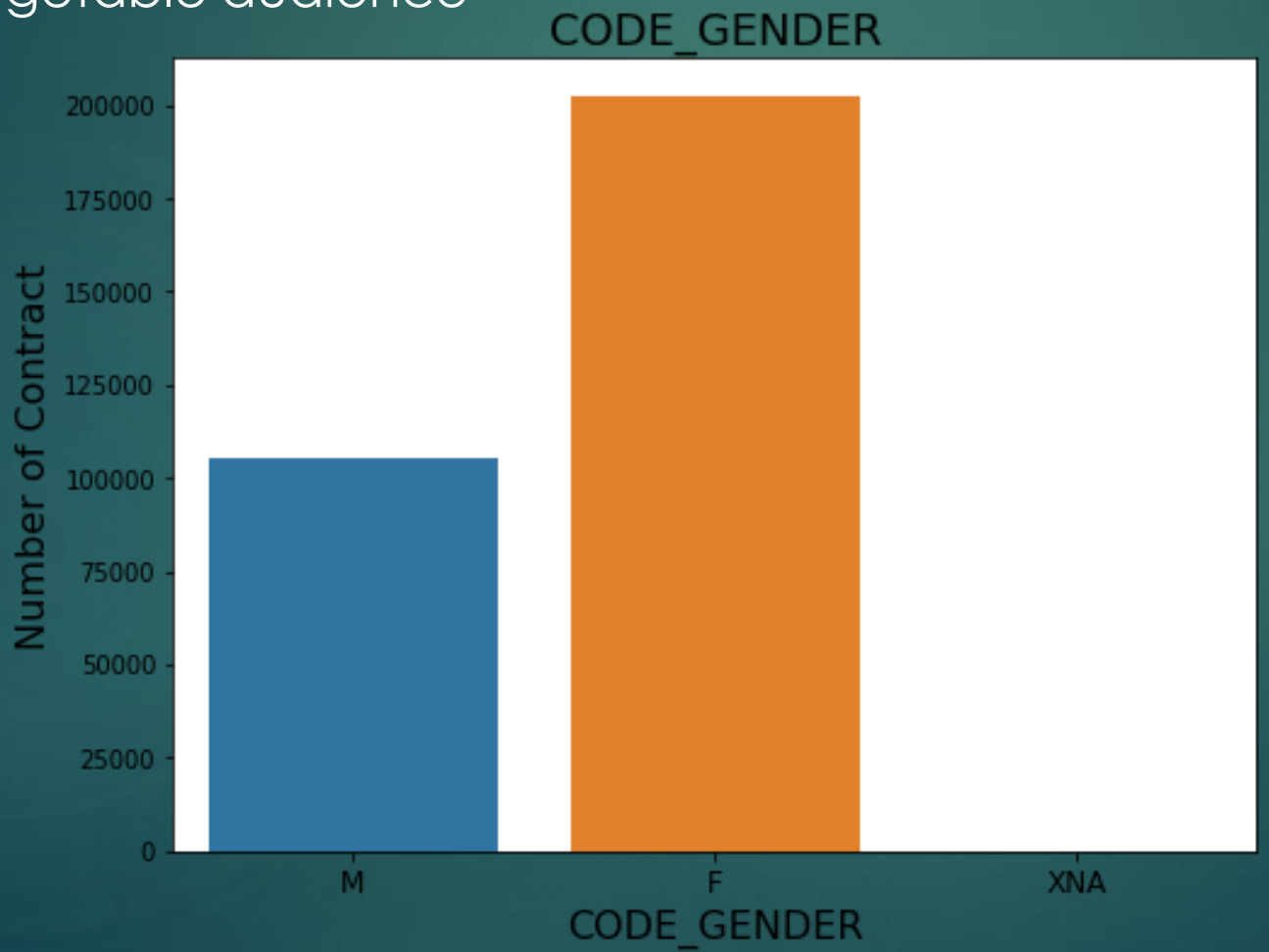
Variable 1 - NAME_CONTRACT_TYPE

- No of cash loans > No of Revolving Loans, shows cash loan attracts better business



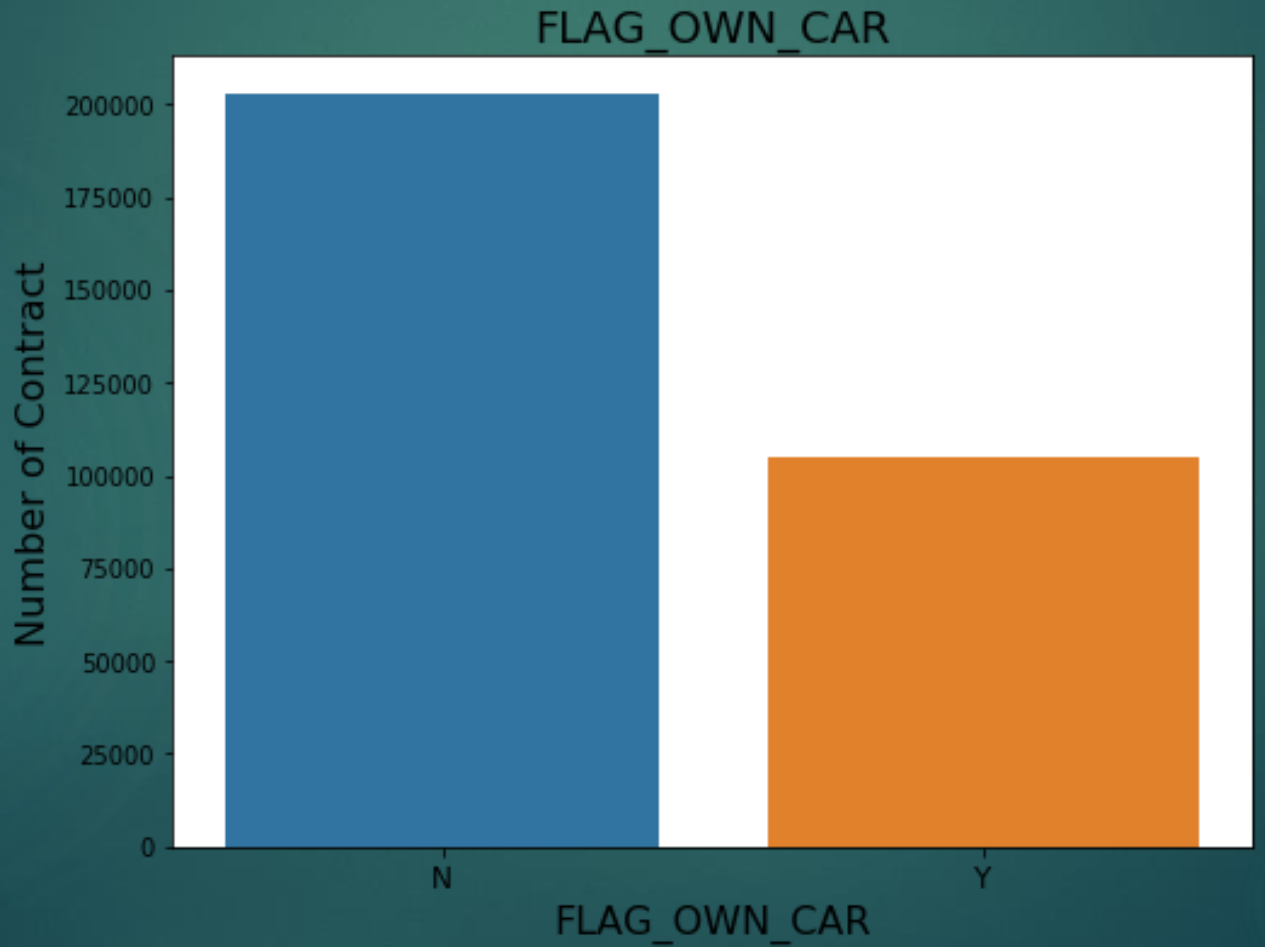
Variable 2 – CODE_GENDER

- No of Female Clients > No of Male Clients, showing female are more targetable audience



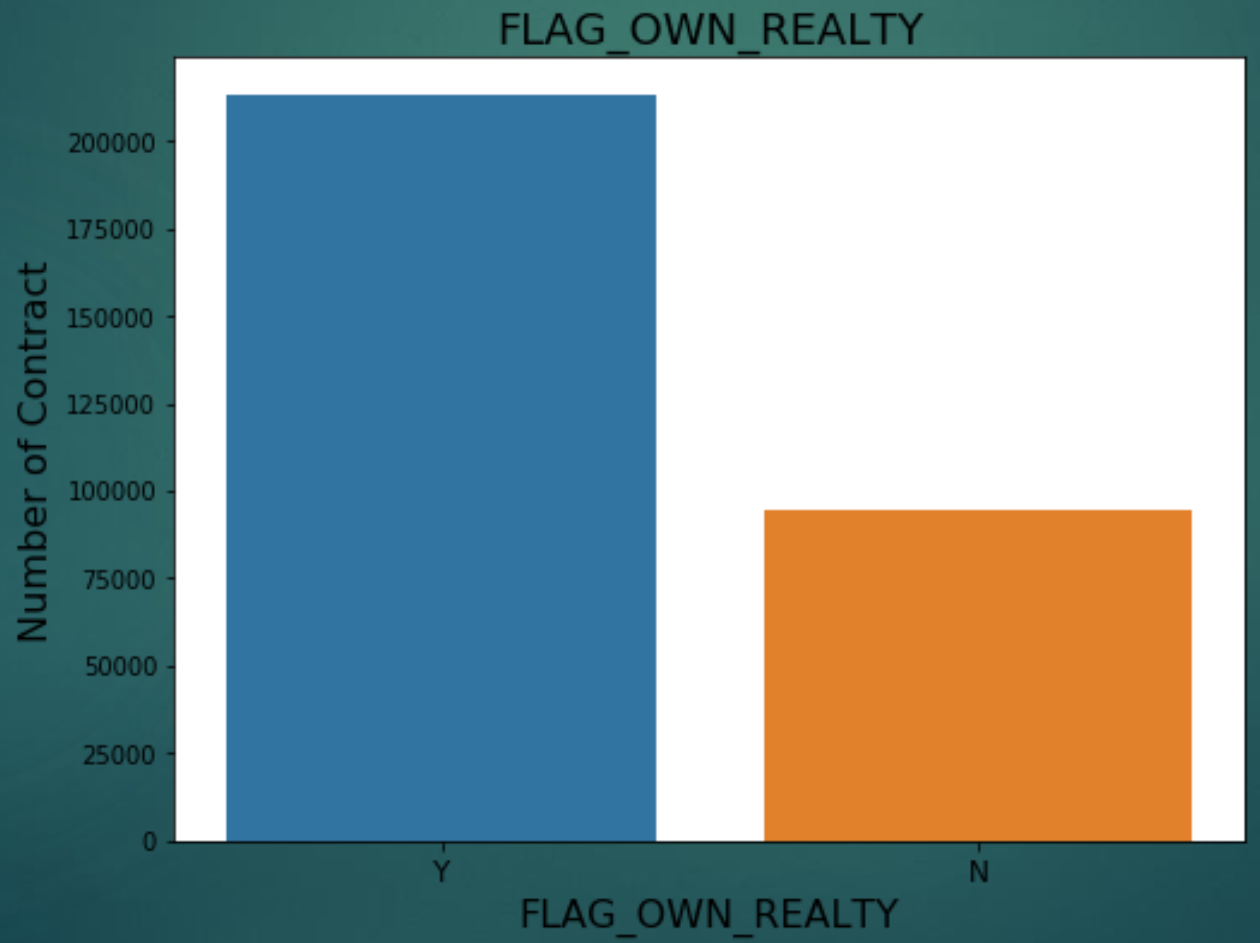
Variable 3 – FLAG_OWN_CAR

- Clients who haven't purchased a car are more likely to fall for a loan than the ones already owning!



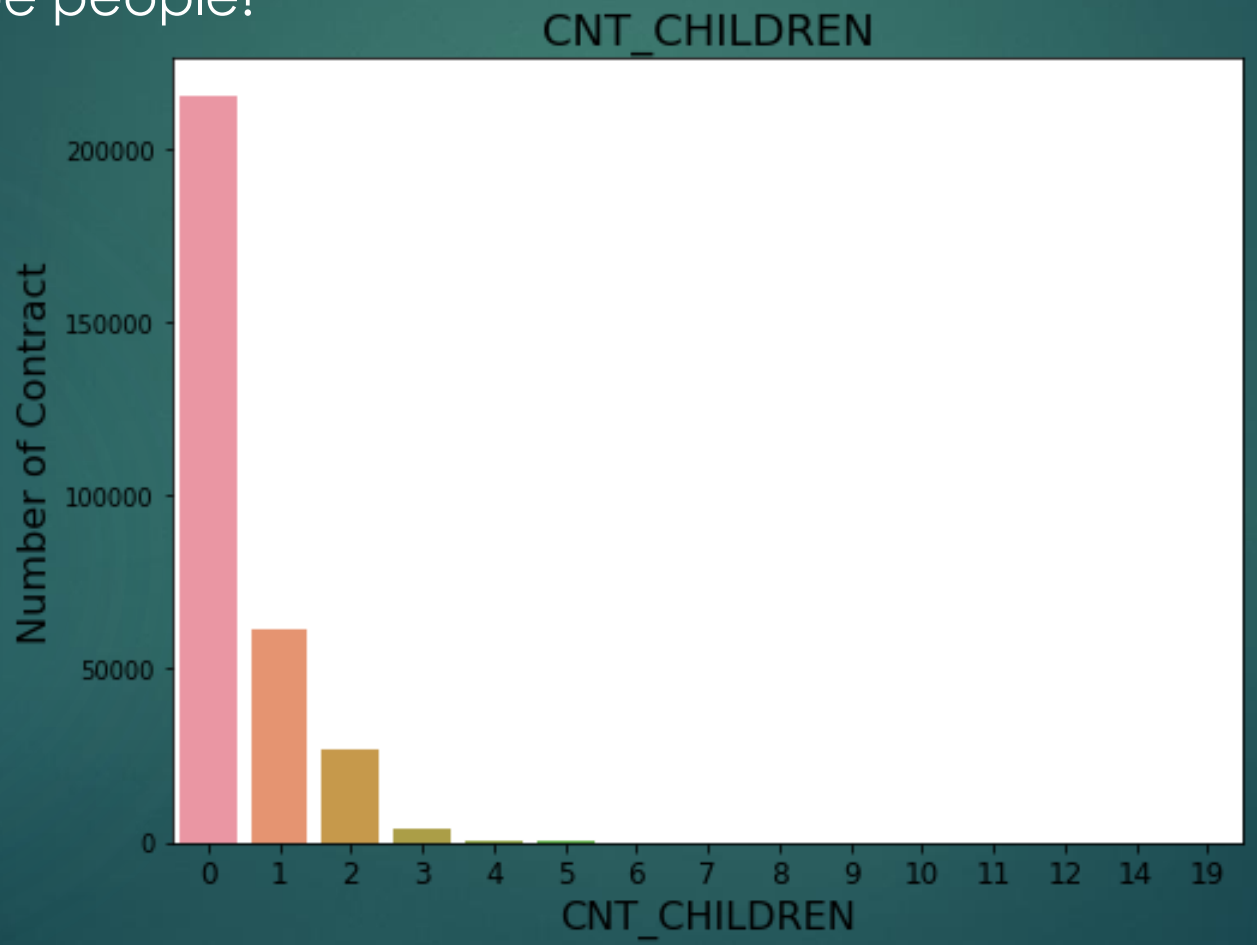
Variable 4 – FLAG_OWN_REALTY

- Clients who are into real estates are more liable and targeted customers than the clients who aren't into it



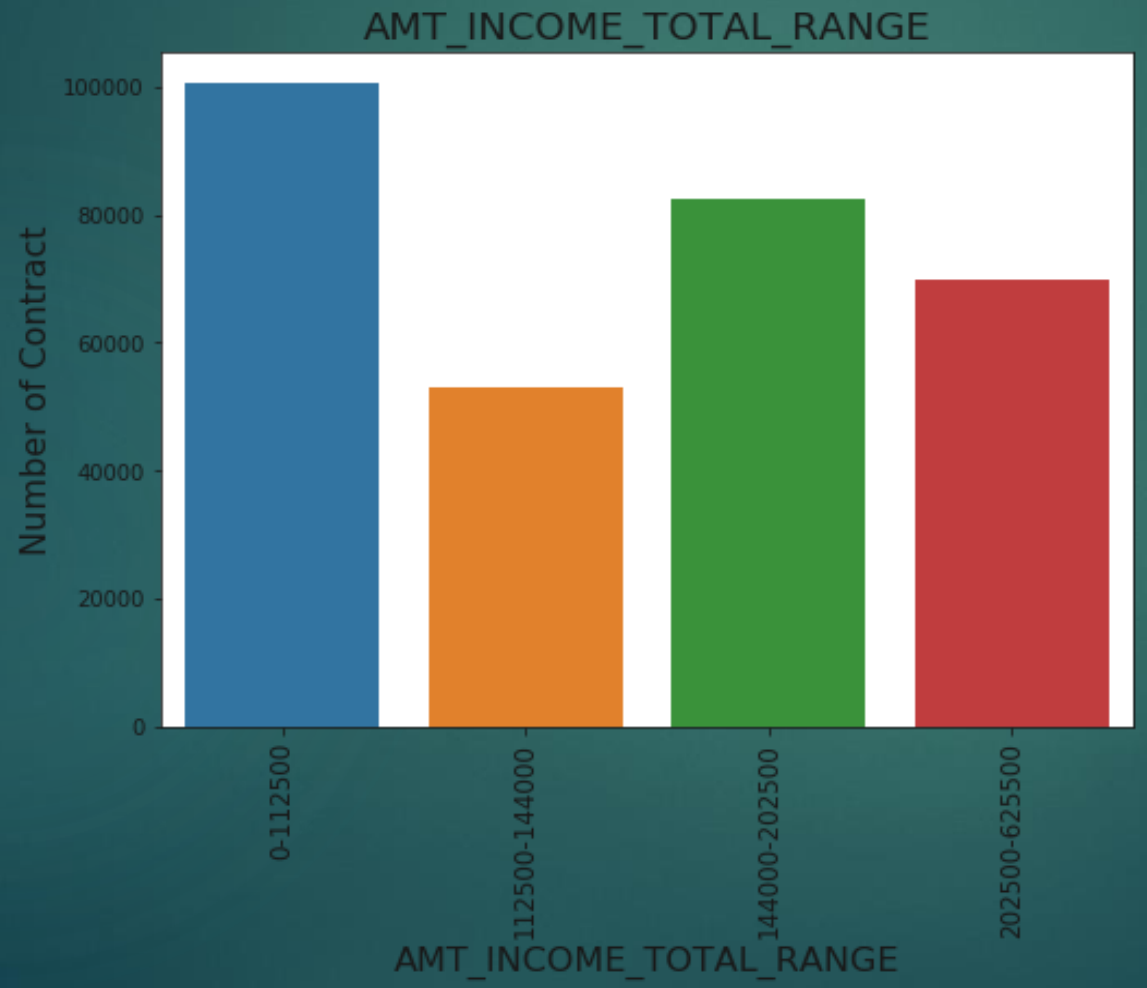
Variable 5 – CNT_CHILDREN

- Clients who aren't parents yet subject themselves to loans rather than family type people!



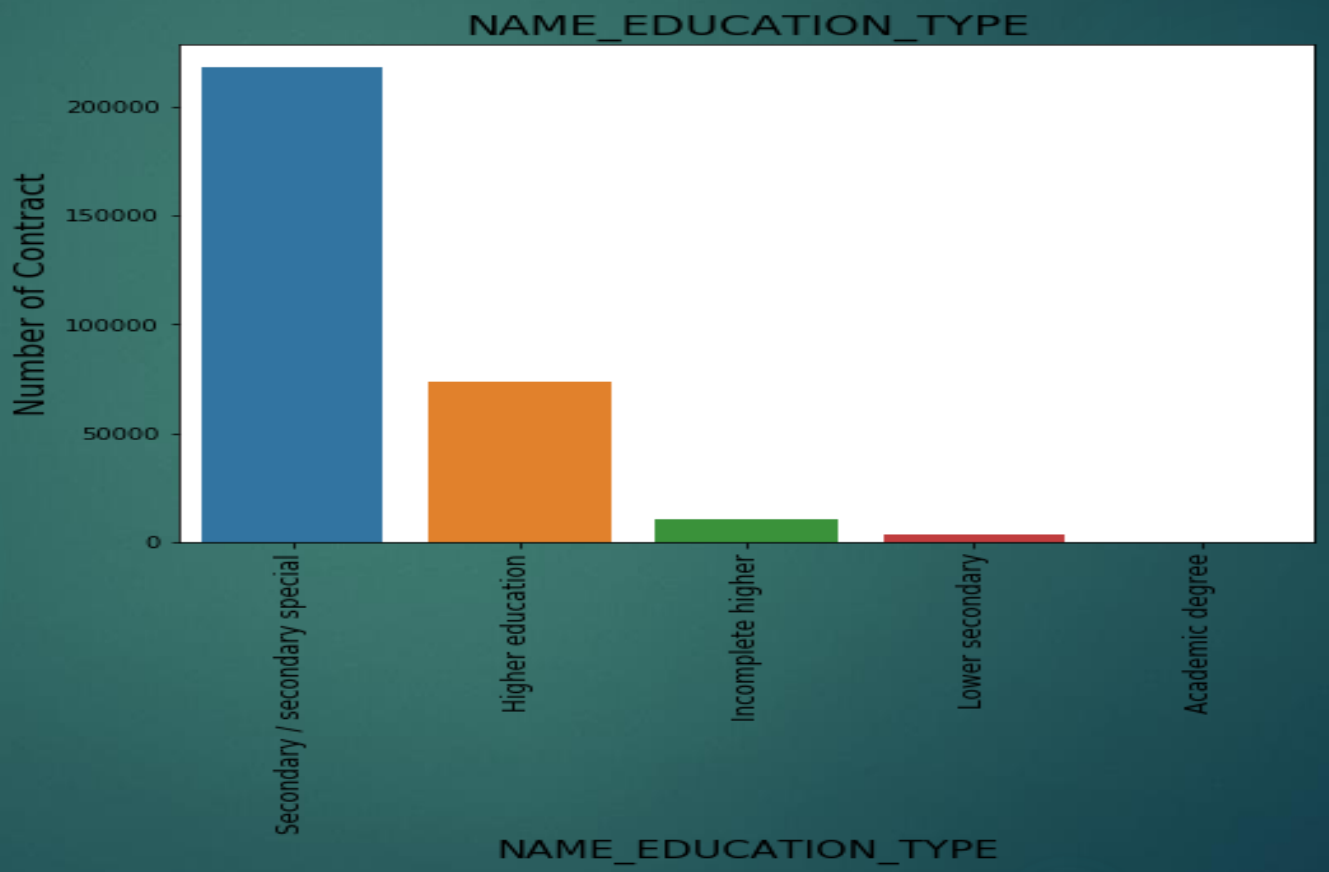
Variable 6 – AMT_INCOME_TOTAL_RANGE

- Most of the clients availing loan have a salary less than 112500



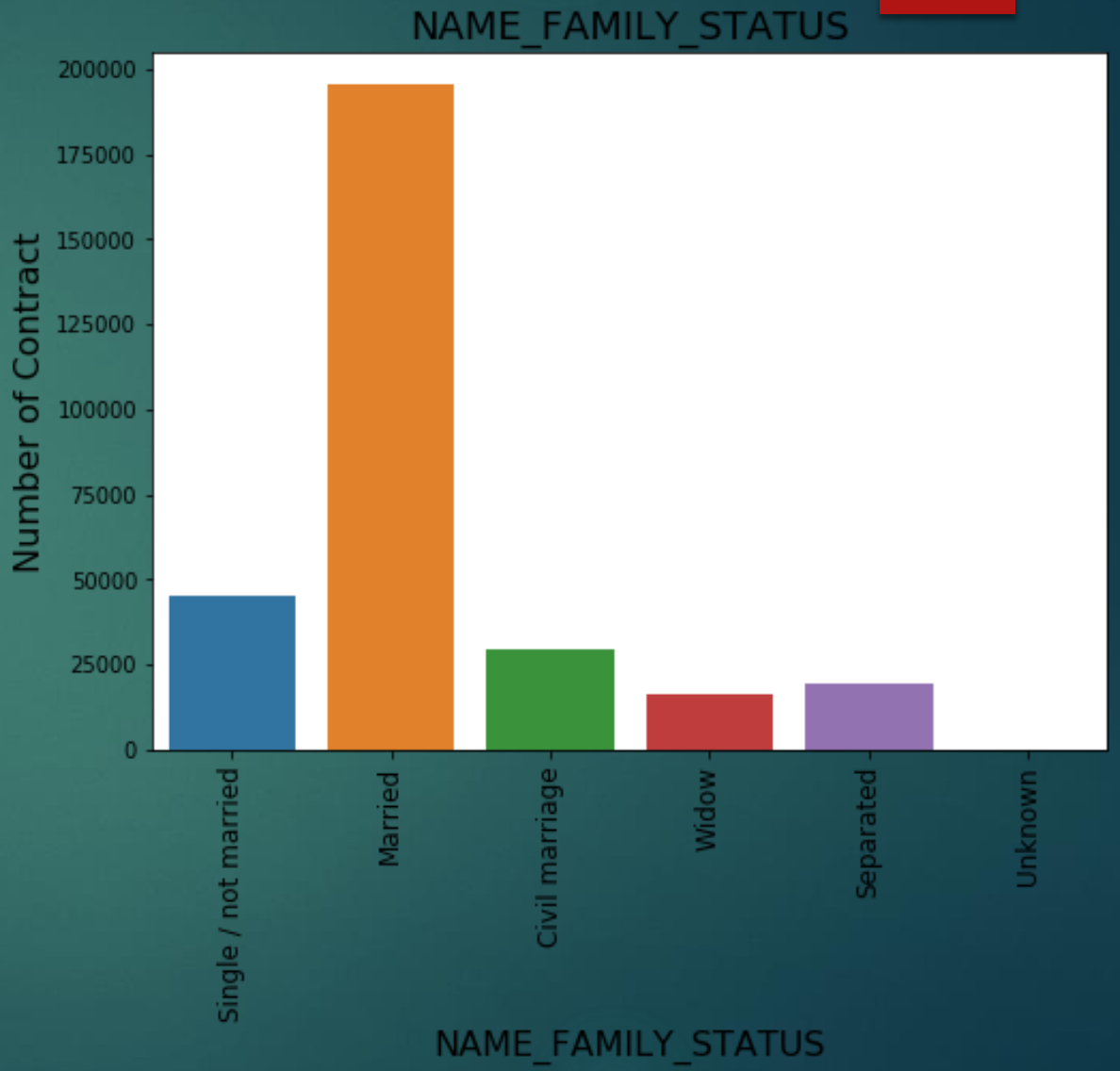
Variable 7 – NAME_EDUCATION_TYPE

- Clients with Education level as secondary depend more on aids than educated groups



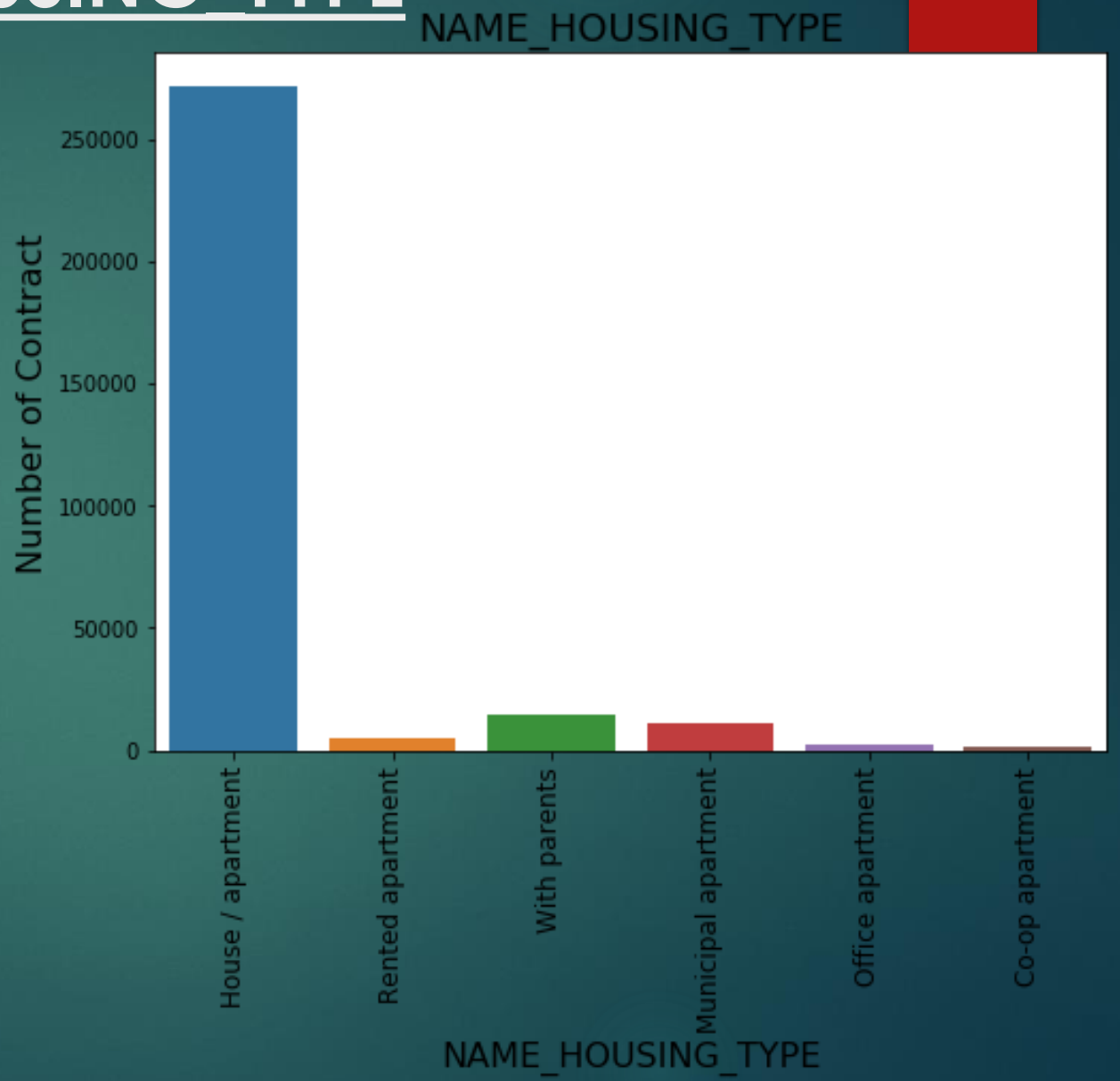
Variable 8 – NAME_FAMILY_STATUS

- Married people tend to take more loans than the other kind of people



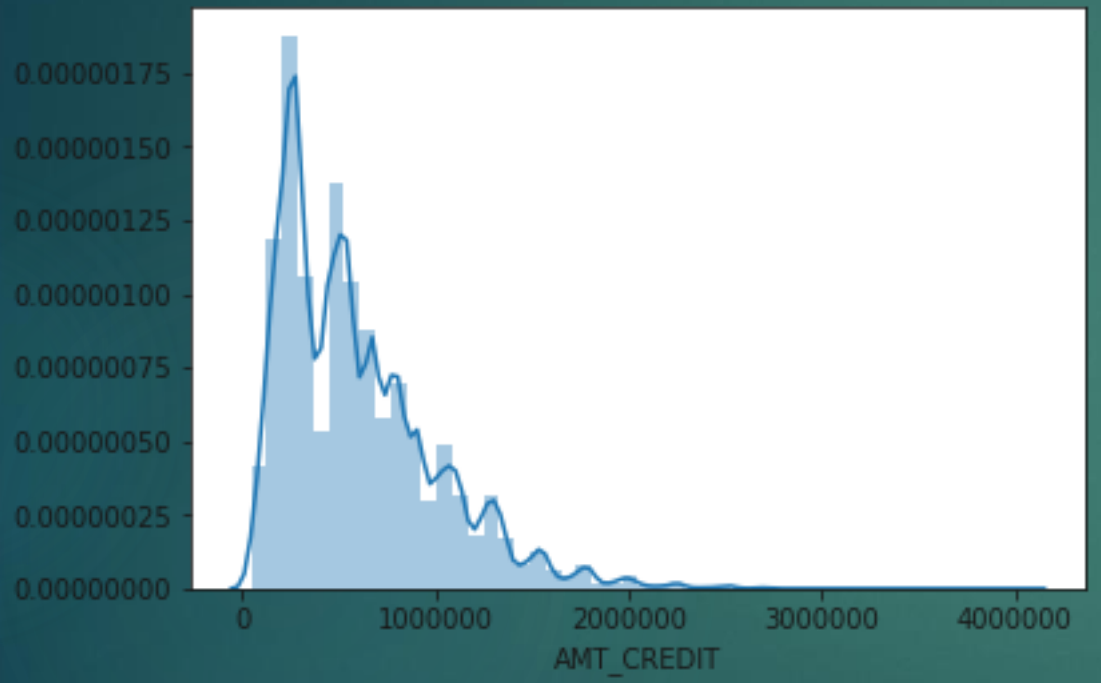
Variable 9 – NAME_HOUSING_TYPE

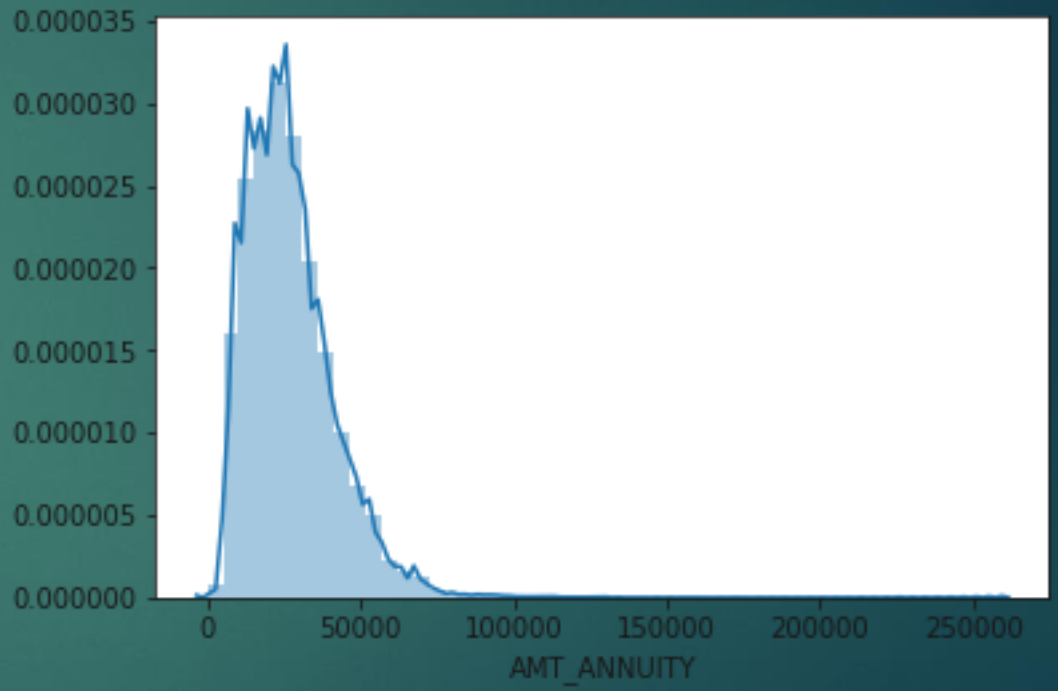
- Person who own House/apartment tends to take more loan than other clients



Variable 10 – Credit Amount

- Credit VS Annuity

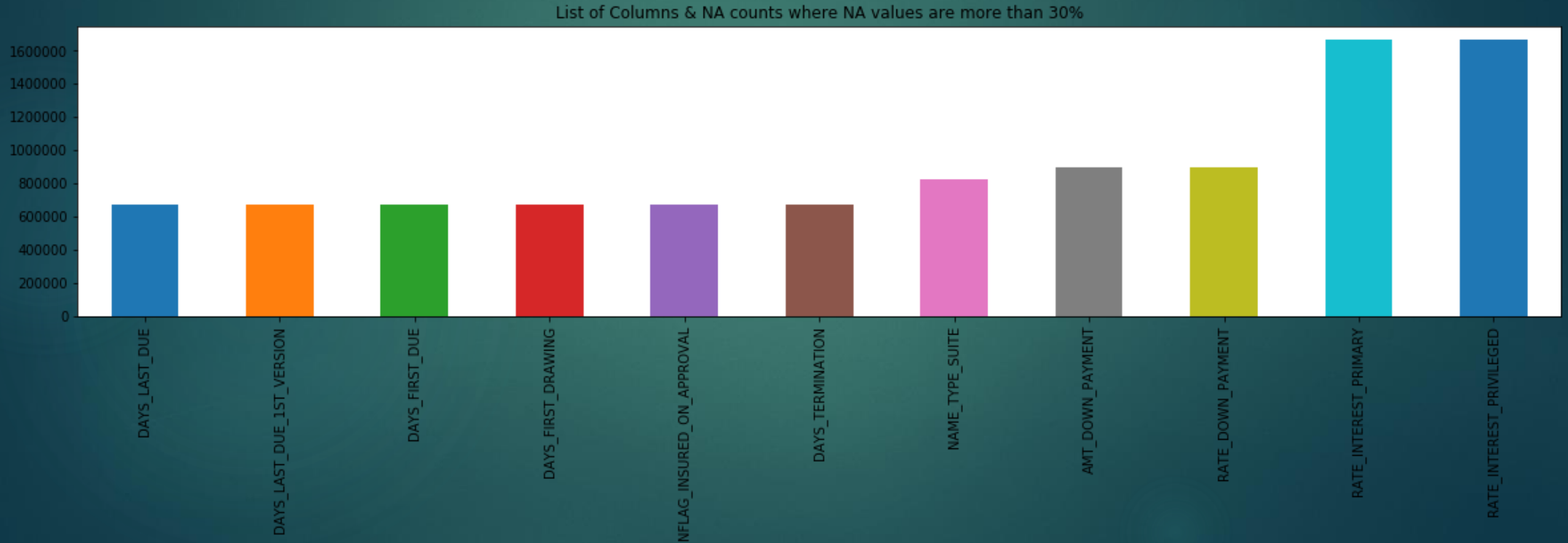




Data Set - Analysis
Previous Application Dataset

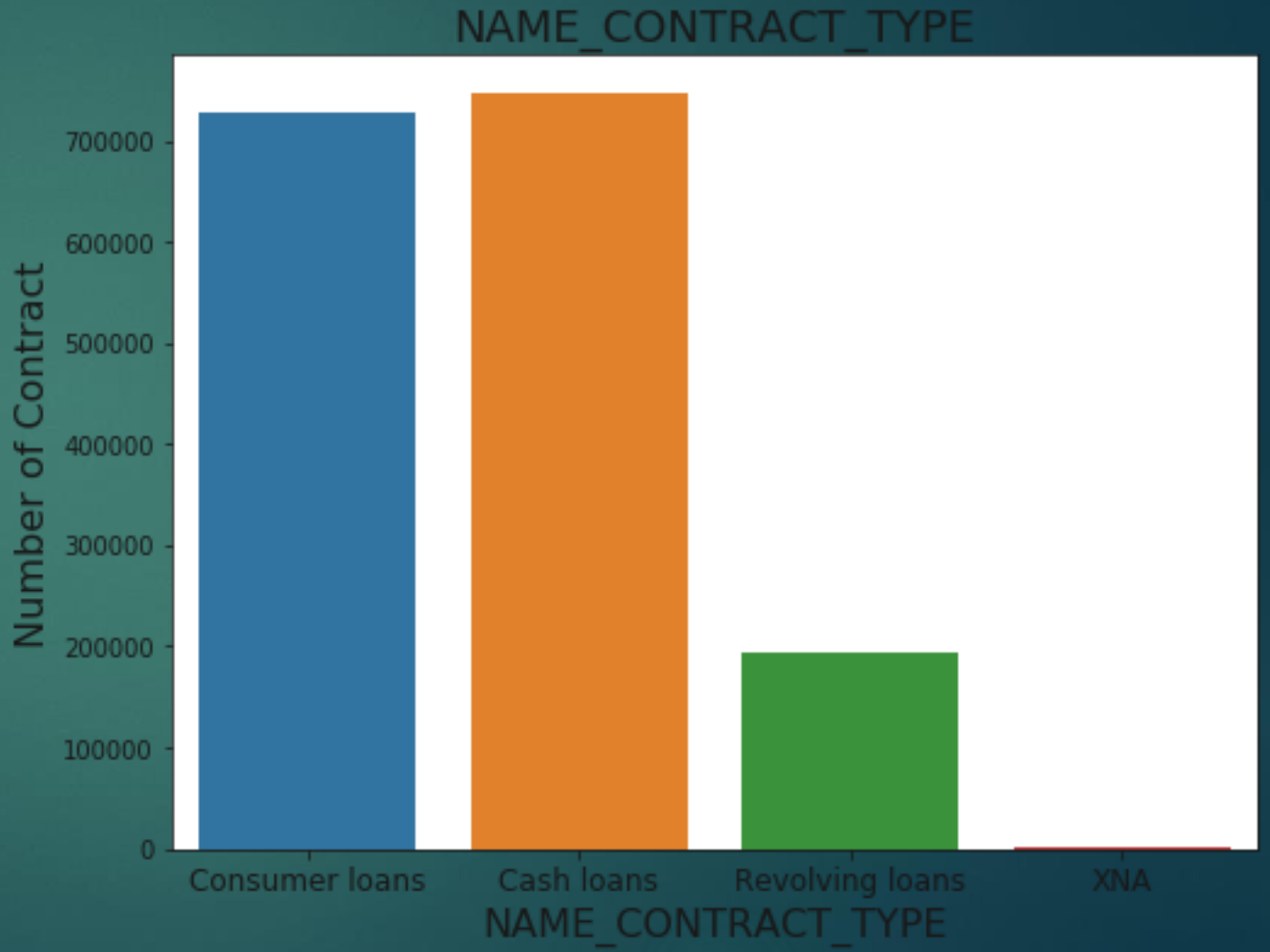
Data Exploration - Previous Application Dataset

- Checking for data columns with less than 30% missing data
- Columns or variables with more than 30% missing data can be removed, since these variables would not account to analysis



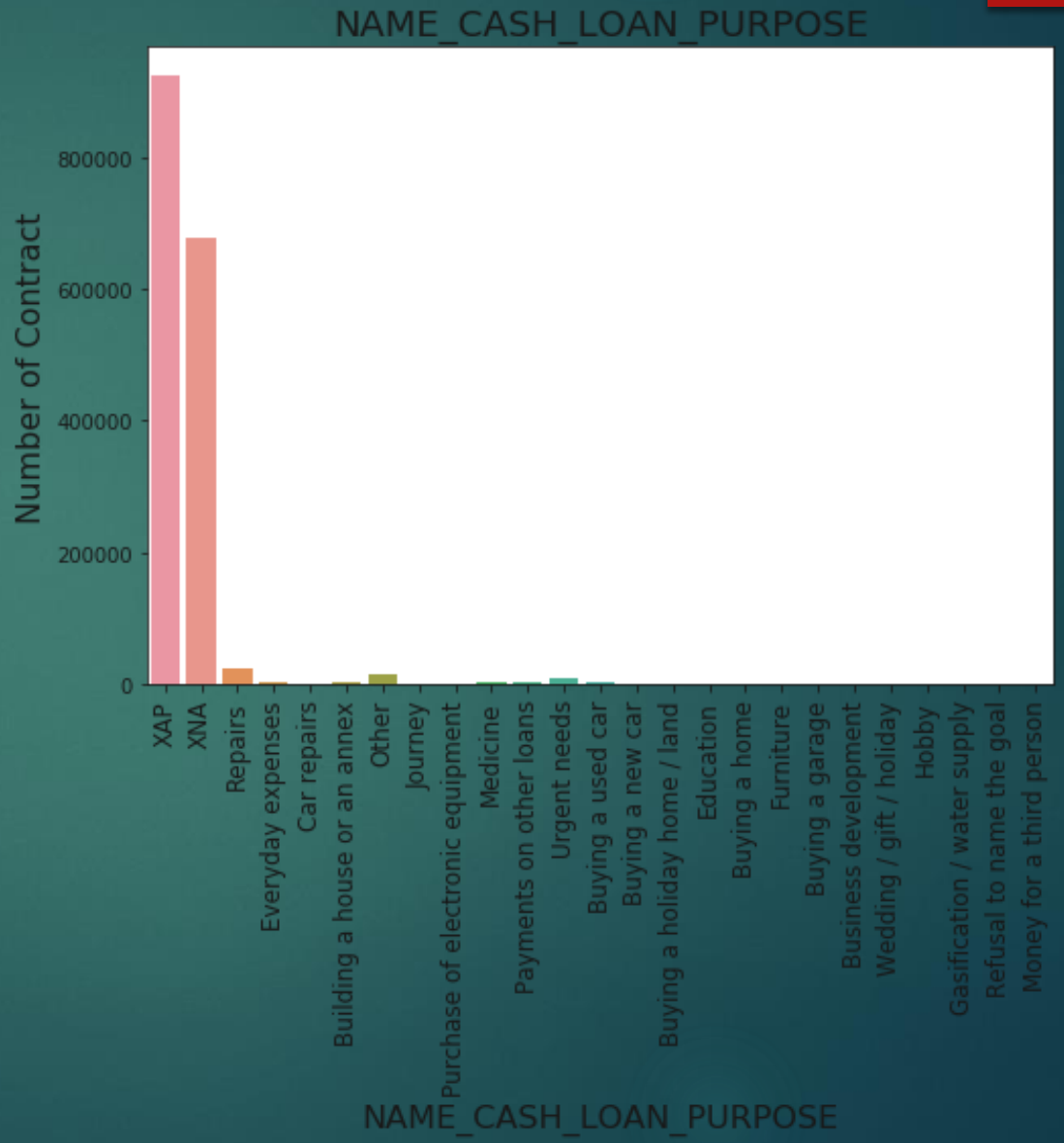
Variable 1 – NAME_CONTRACT_TYPE

- Consumer and cash loans are the major source of loan type and almost of the same demands



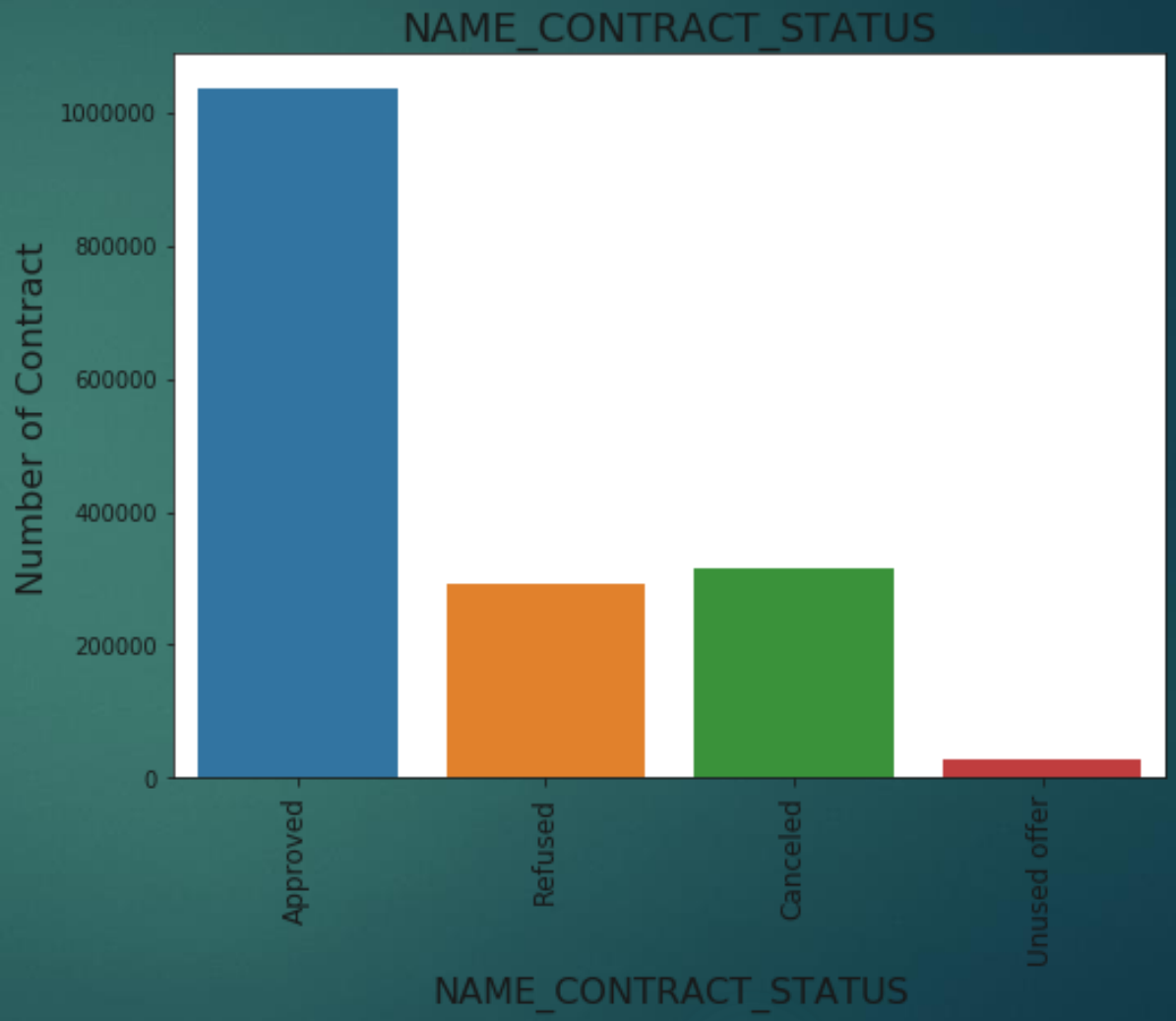
Variable 2 – CASH_LOAN_PURPOSE

- Urgent needs, medicine, building a house or an annex accounts for the largest number of contracts.



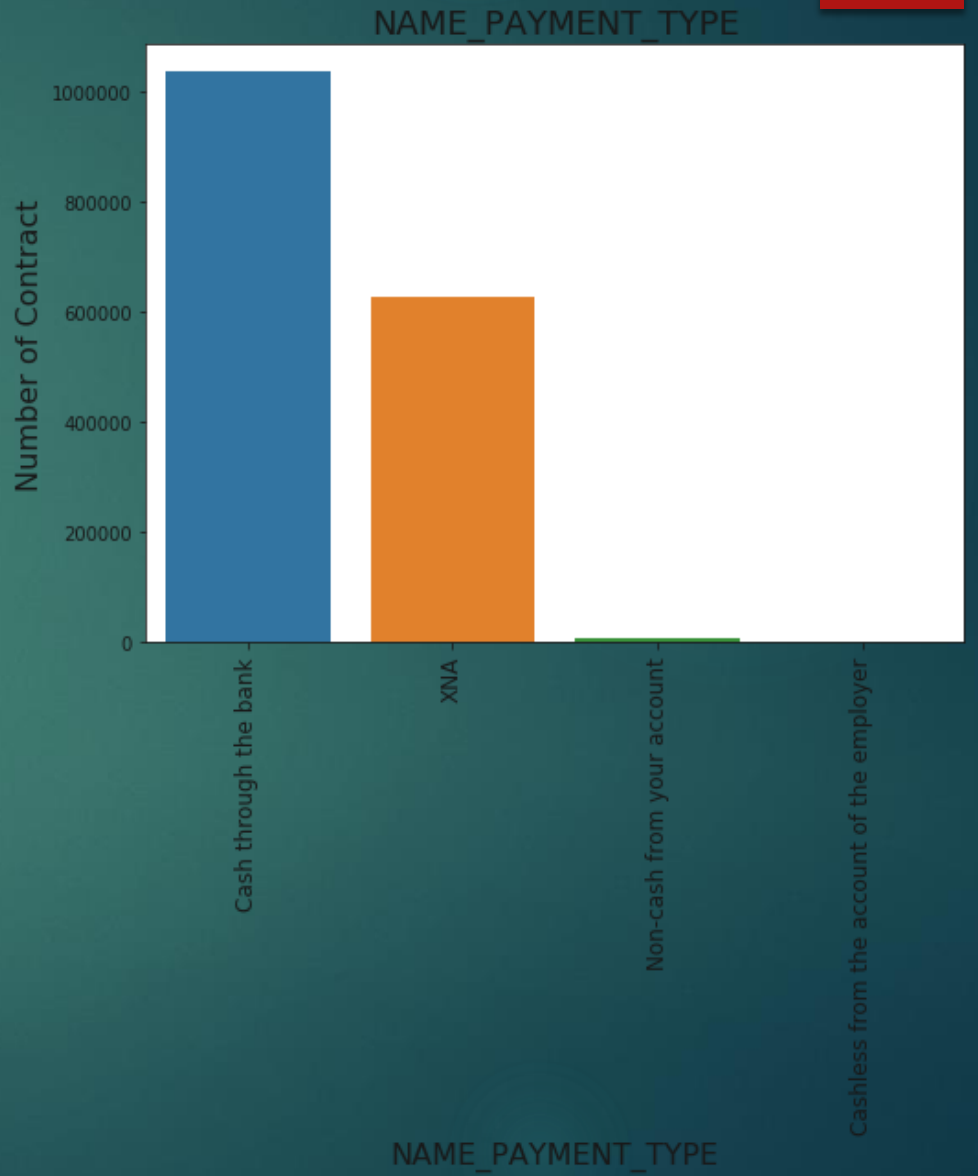
Variable 3 – NAME_CONTRACT_STATUS

- Higher percentage of previous loans have been approved rather than being refused or cancelled



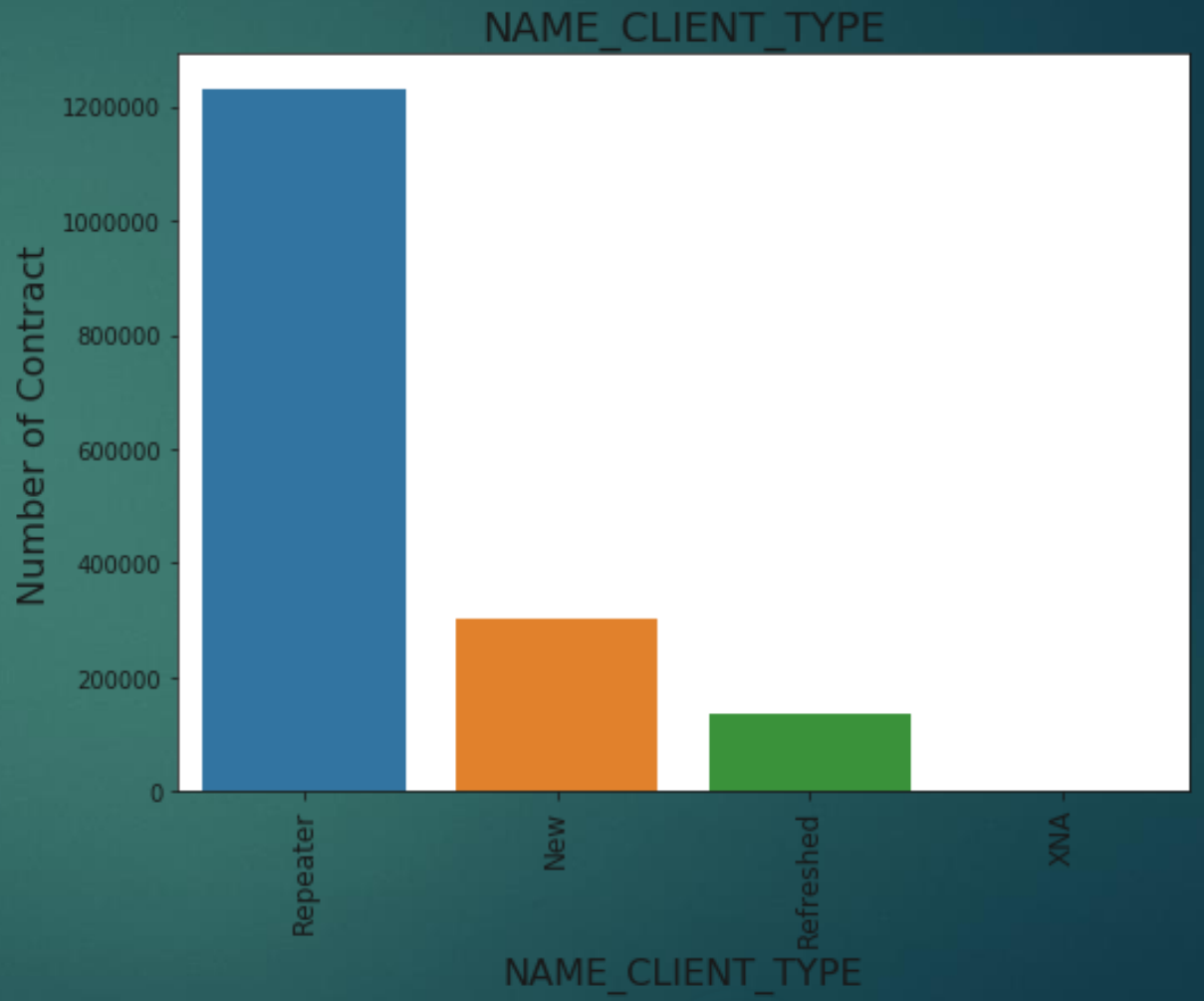
Variable 4 – NAME_PAYMENT_TYPE

- Cash through Bank has been the major source of payments for the previous applicants



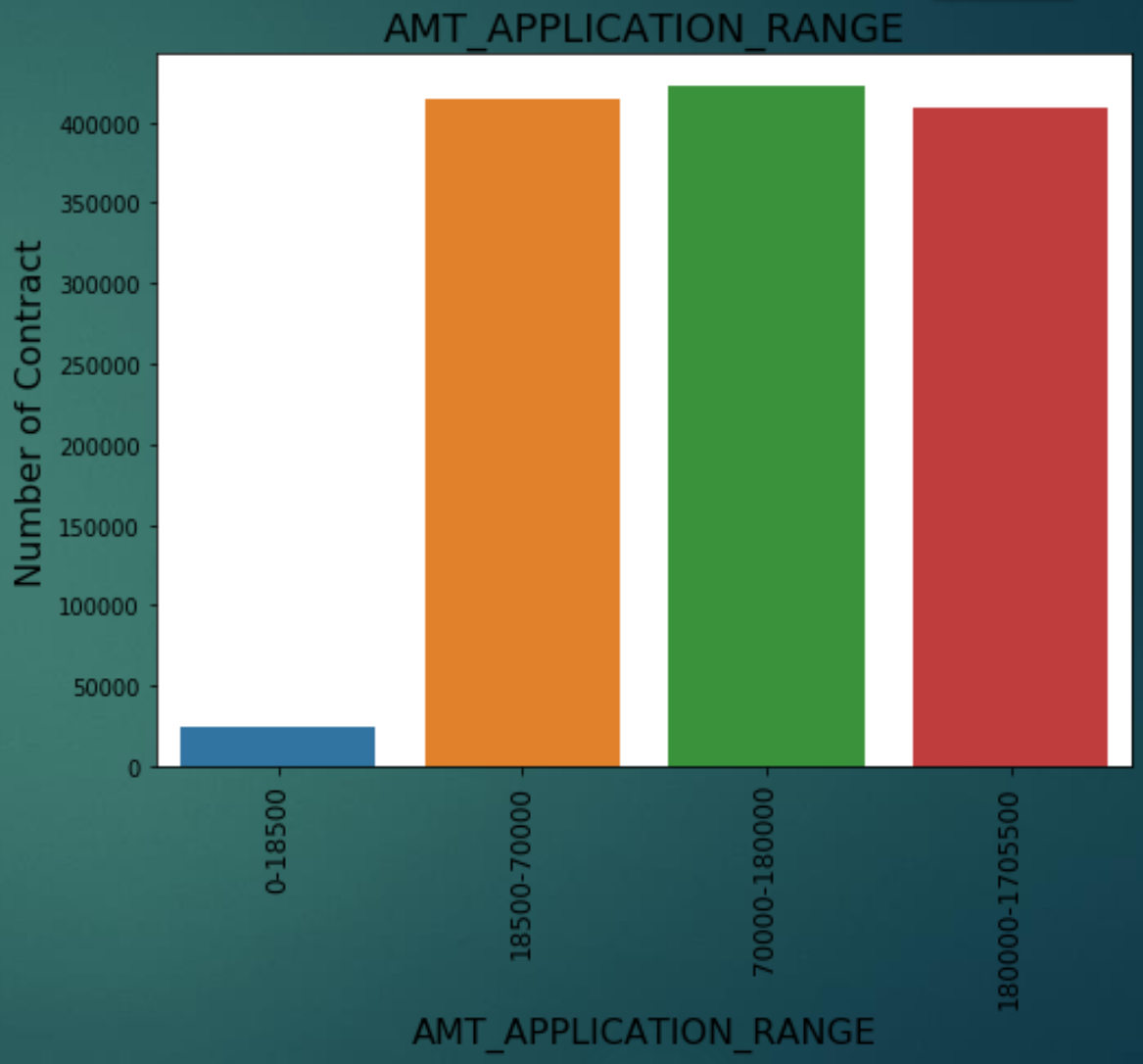
Variable 5 – NAME_CLIENT_TYPE

- Customers who have applied or availed for a loan earlier are more likely to take future loans



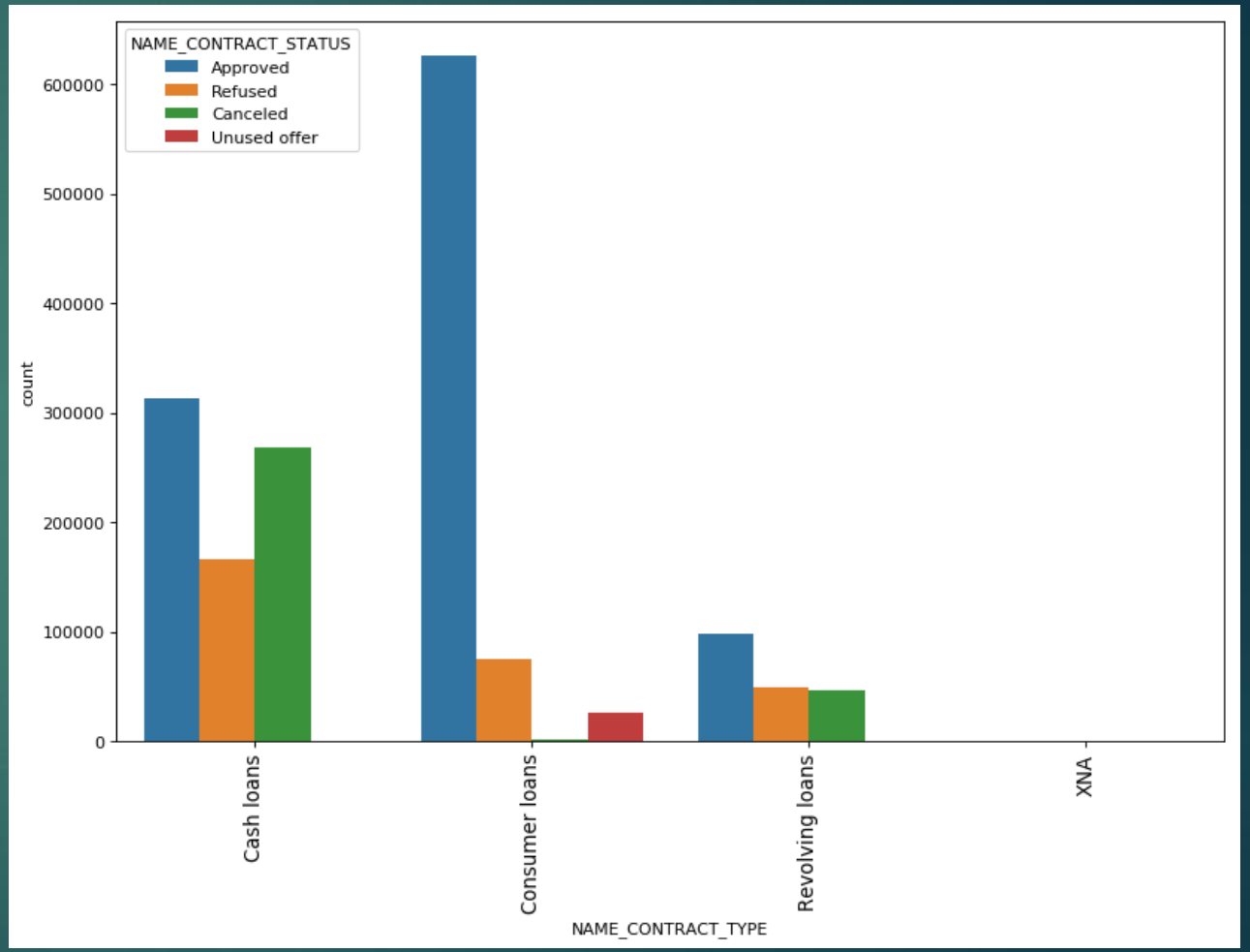
Variable 6 – AMT_APPLICATION_RANGE

- Loans availed are most likely above amount of Rs.18500



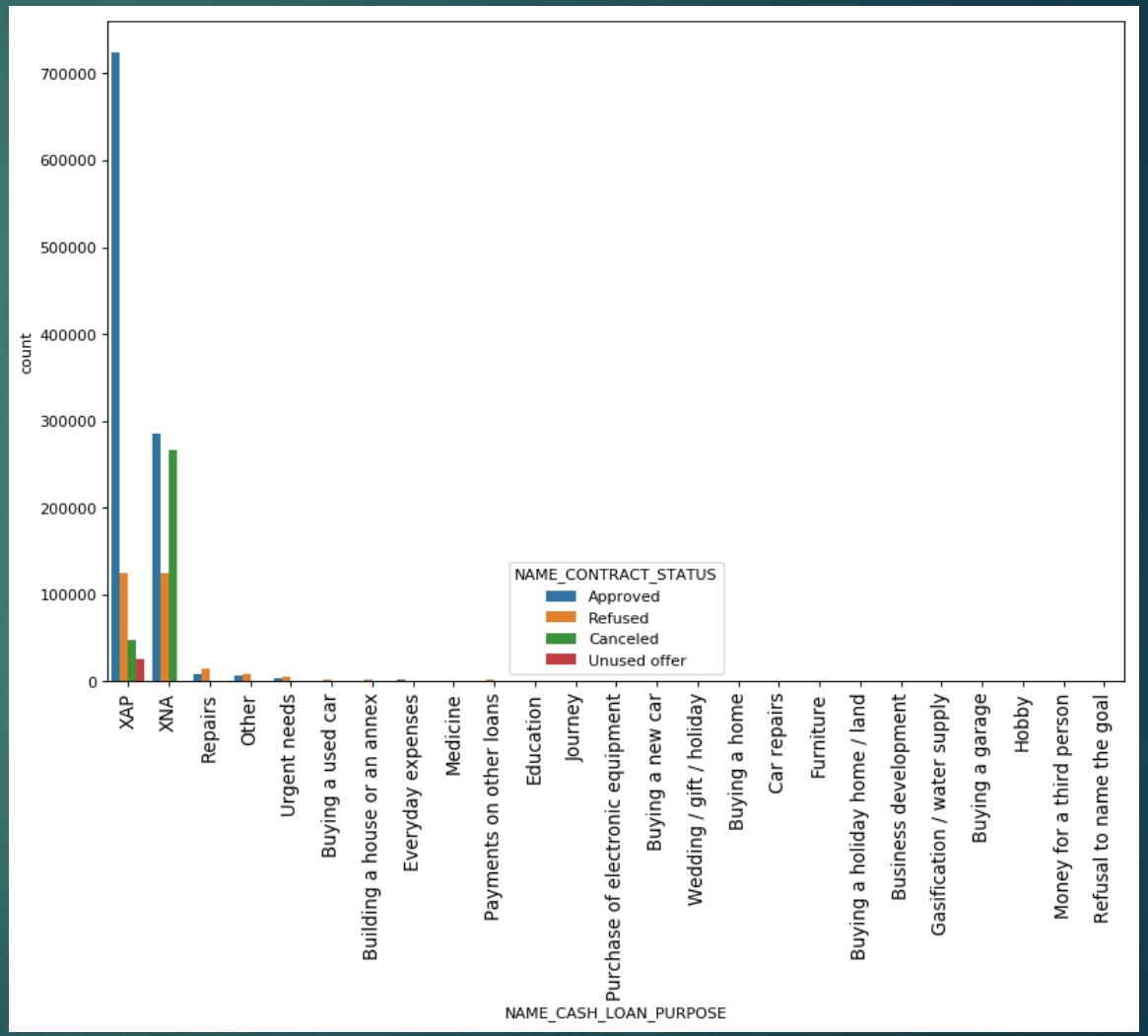
Variable 7 – Bivariate Type and Status

- Most of the consumer loans approved are of type - Cash



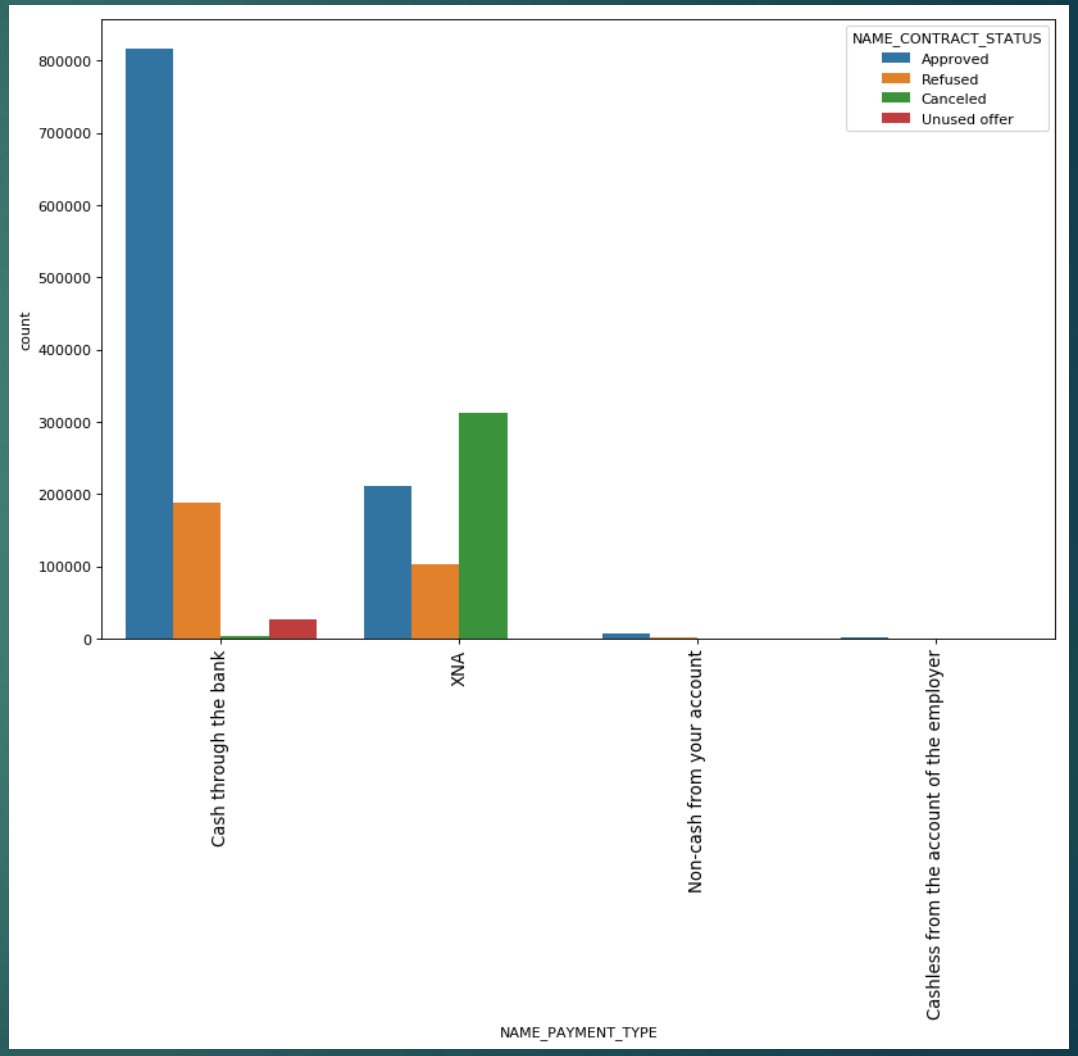
Variable 8 – Bivariate Loan Purpose and Status

- Majority of the loans approved do not hold a strong reason or purpose



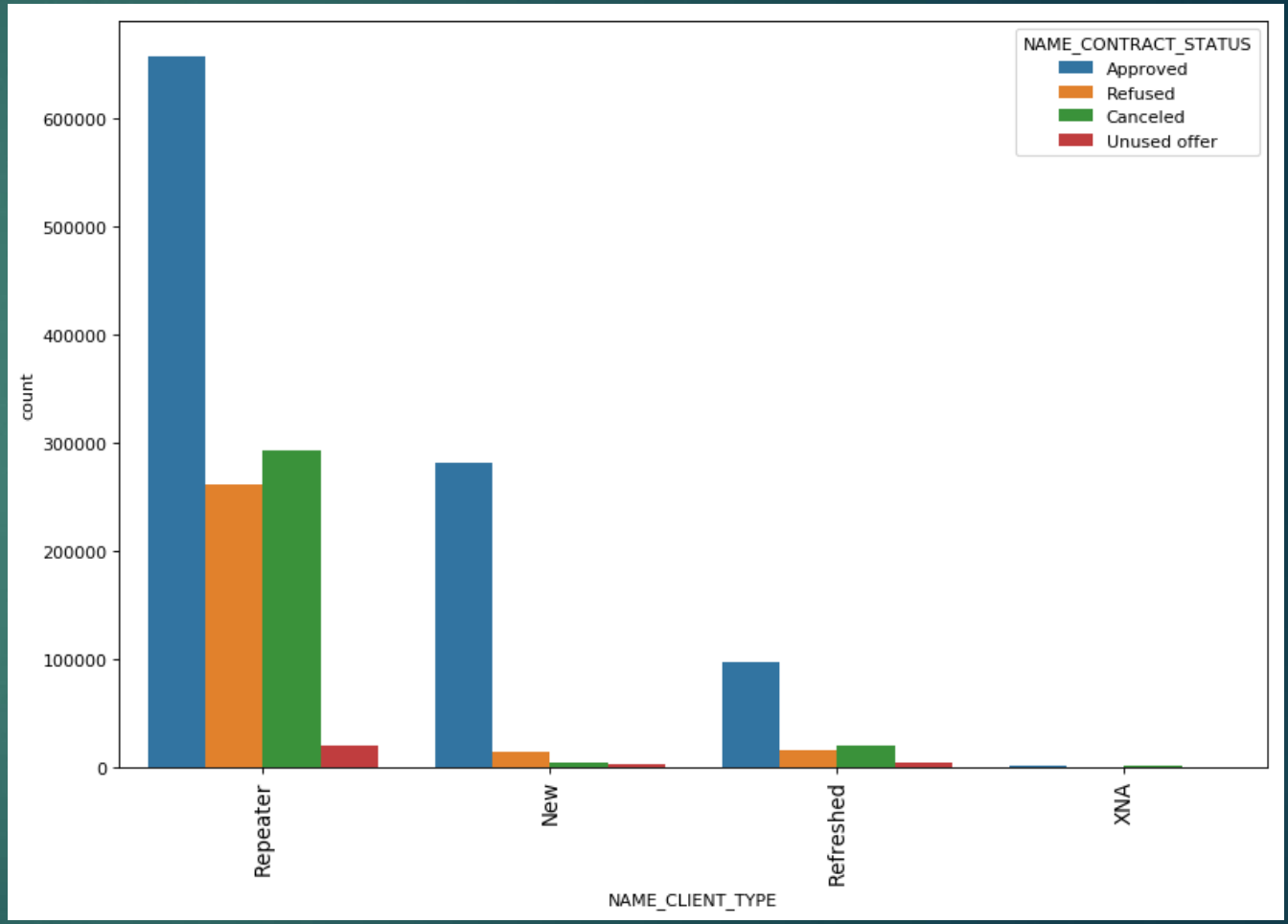
Variable 9 – Bivariate Payment Type and Status

- Majority of the loans approved are in the form of cash through bank payment mode



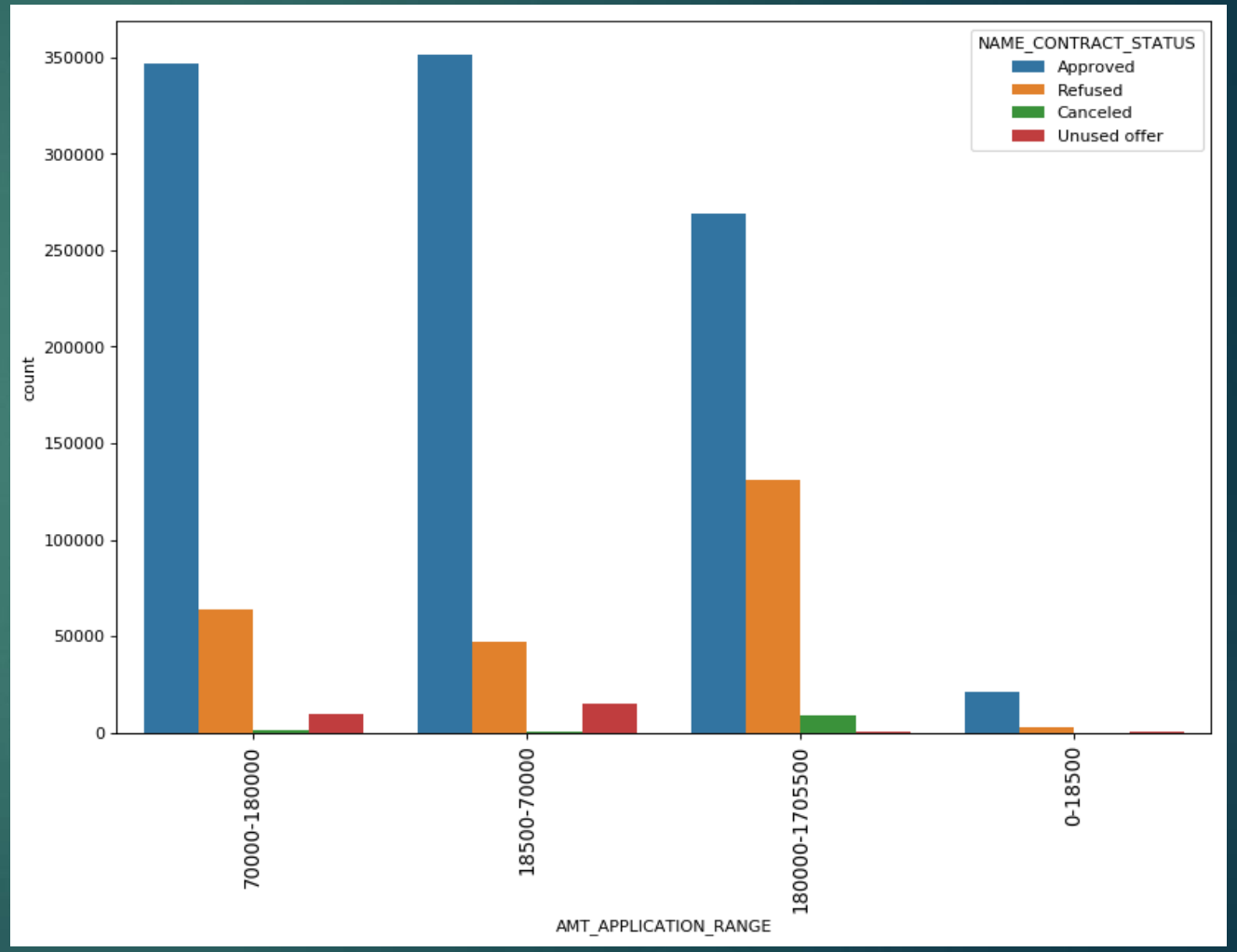
Variable 10 – Bivariate Client Type and Status

- Majority of the loans approved are the ones requested by a new customer
- This shows a low customer retention rate



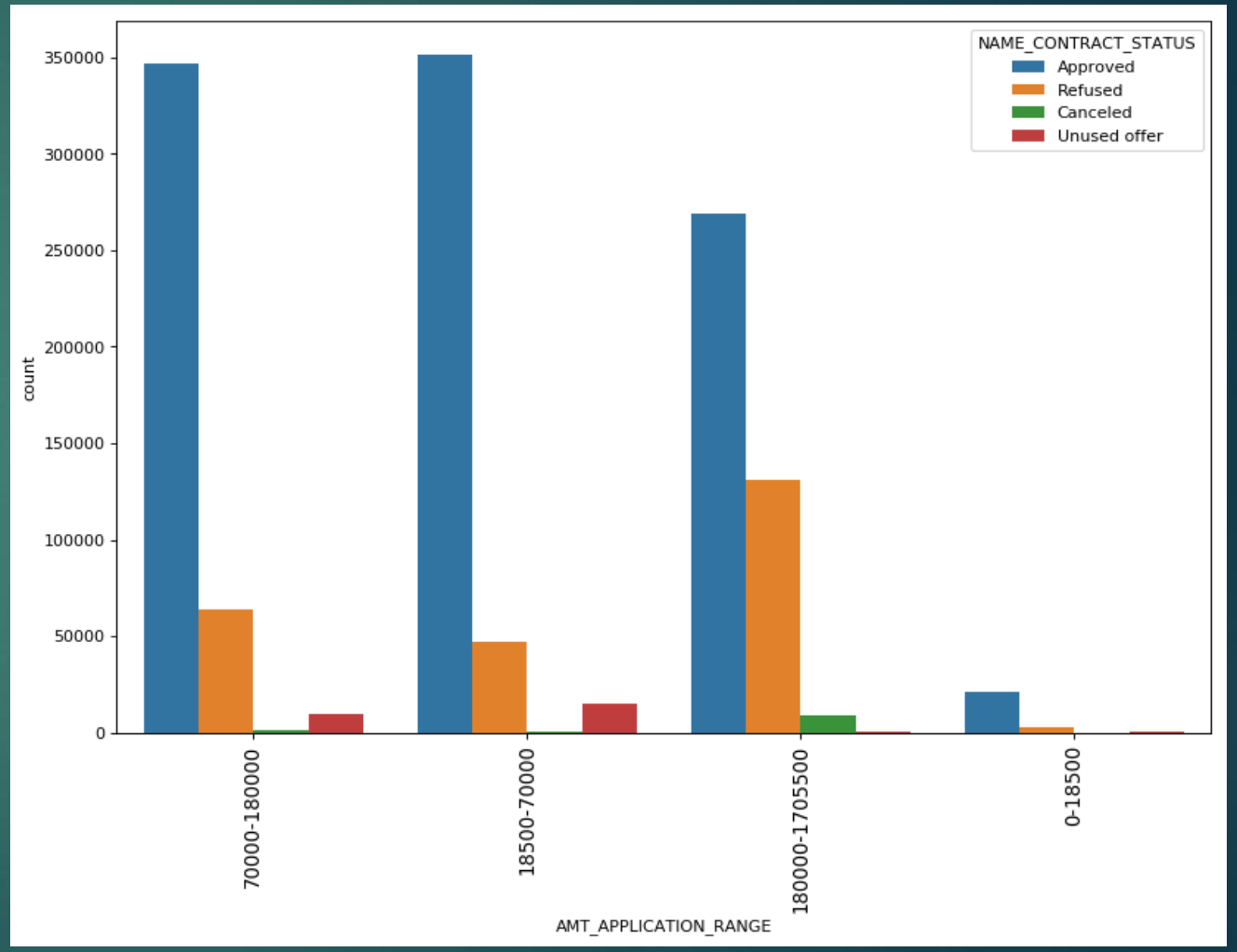
Variable 11 – Bivariate Amnt Range and Status

- Loans with range in between 70000-180000 and 18500-70000 has more approve rate compare to that of 180000-1705500 range.



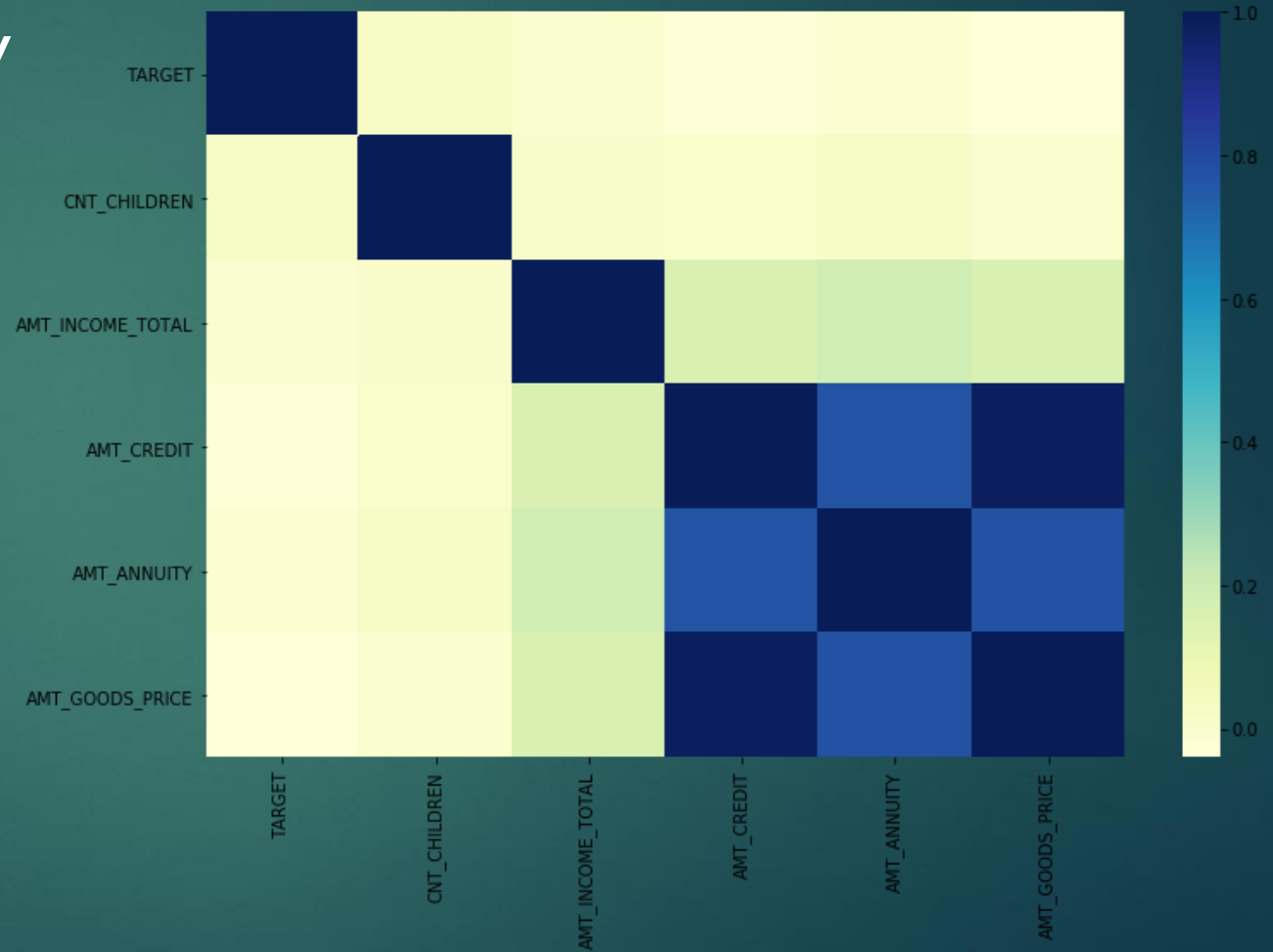
Variable 12 – Bivariate Amnt Range and Status

- Loans with range in between 70000-180000 and 18500-70000 has more approve rate compare to that of 180000-1705500 range.



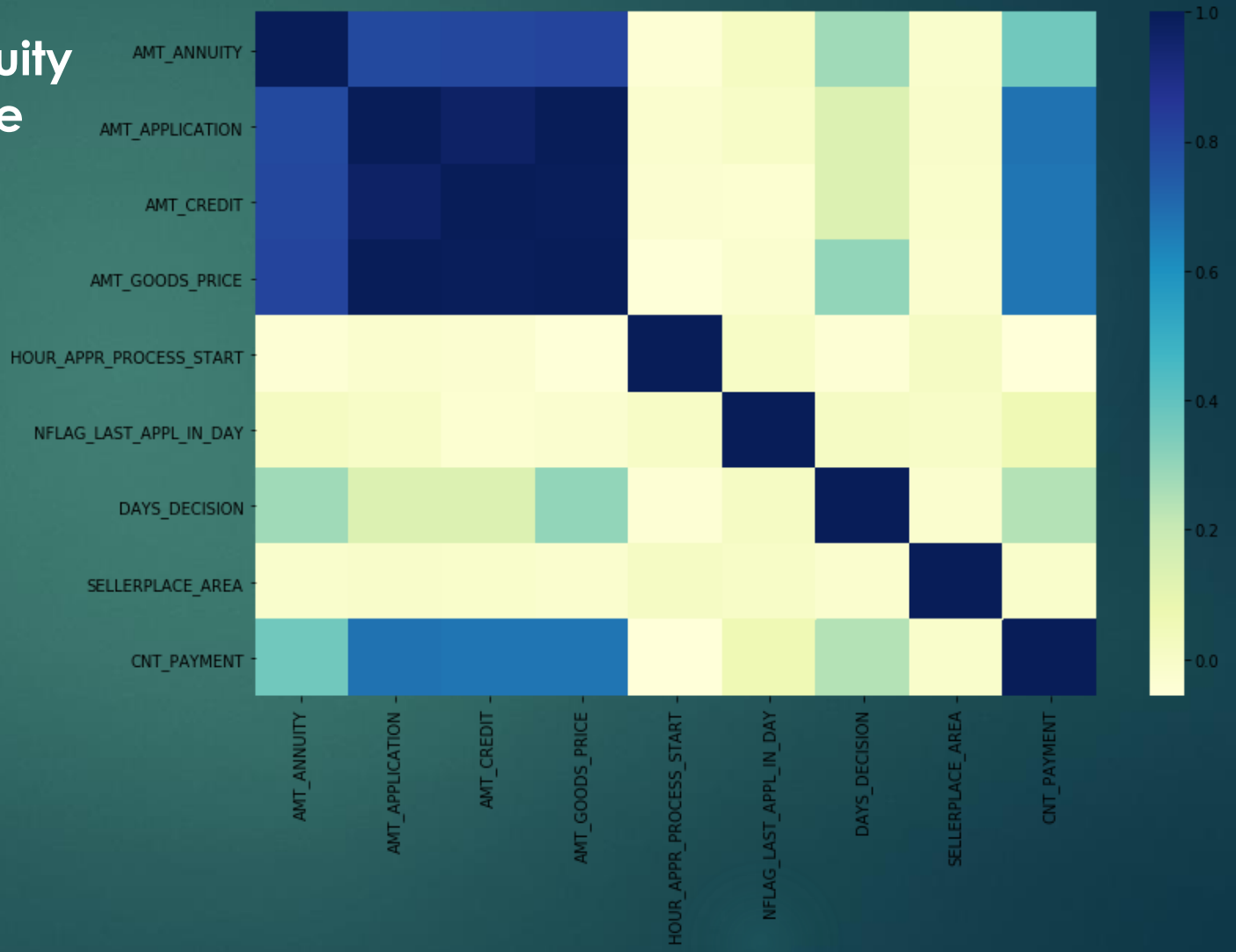
Heat Map of all variables for Application Dataset

- Amount credit, Amount annuity and Amount Goods Price are highly interrelated



Heat Map of all variables for Previous Application Dataset

- Amount credit, Amount annuity and Amount Goods Price are highly interrelated



Outlier Analysis

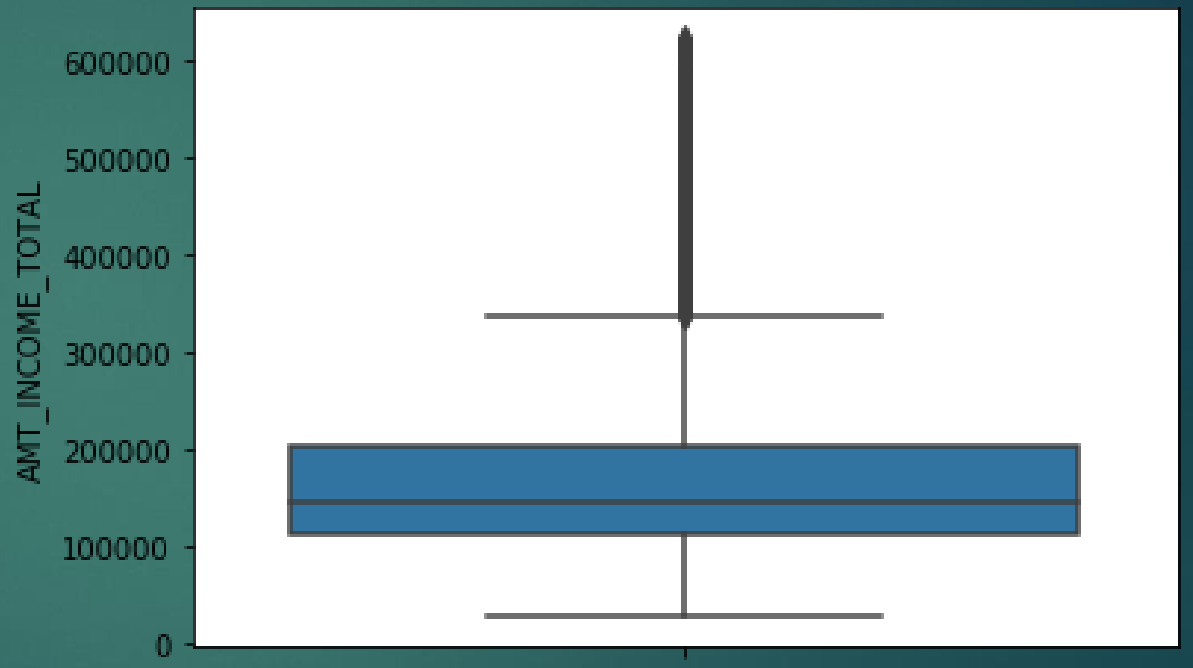
Application Dataset

Max value is much higher (approx~550 times) compare to that of the 3rd quartile of the dataset

- AMT_INCOME_TOTAL
- count 307511
- mean 168797.9193
- std 237123.1463
- min 25650.0000
- 25% 112500.0000
- 50% 147150.0000
- 75% 202500.0000
- max 117000000.0000

Removing outliers (values from 99 to 100%)

- AMT_INCOME_TOTAL
- count 305943
- mean 164764.1491
- std 81644.2477
- min 25650.0000
- 25% 112500.0000
- 50% 144000.0000
- 75% 202500.0000
- max 625500.0000



Outlier Analysis

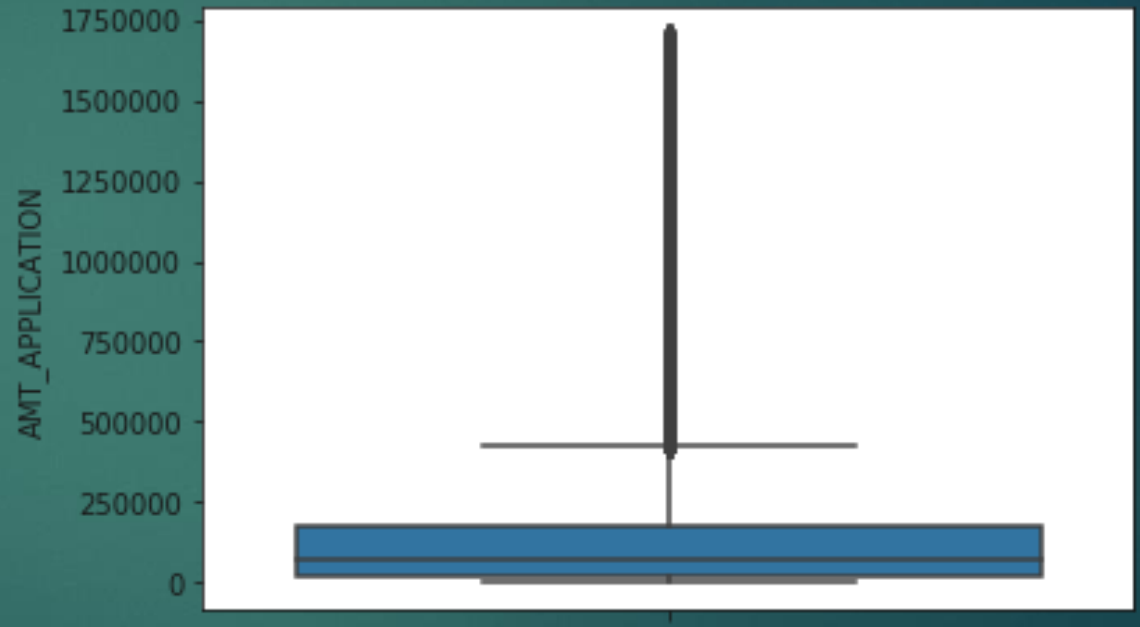
Previous Application Dataset

Maximum loan amount in previous application is about 40 times to that of mean value

- AMT_APPLICATION
- count 1.67021e+06
- mean 175233.8604
- std 292779.7624
- min 0.0000
- 25% 18720.0000
- 50% 71046.0000
- 75% 180360.0000
- max 6905160.0000

Removing outliers (values from 99 to 100%)

- AMT_APPLICATION
- count 1.66169e+06
- mean 165382.0669
- std 257673.1529
- min 0.0000
- 25% 18314.9212
- 50% 70101.0000
- 75% 180000.0000
- max 1705500.0000



Thank You!