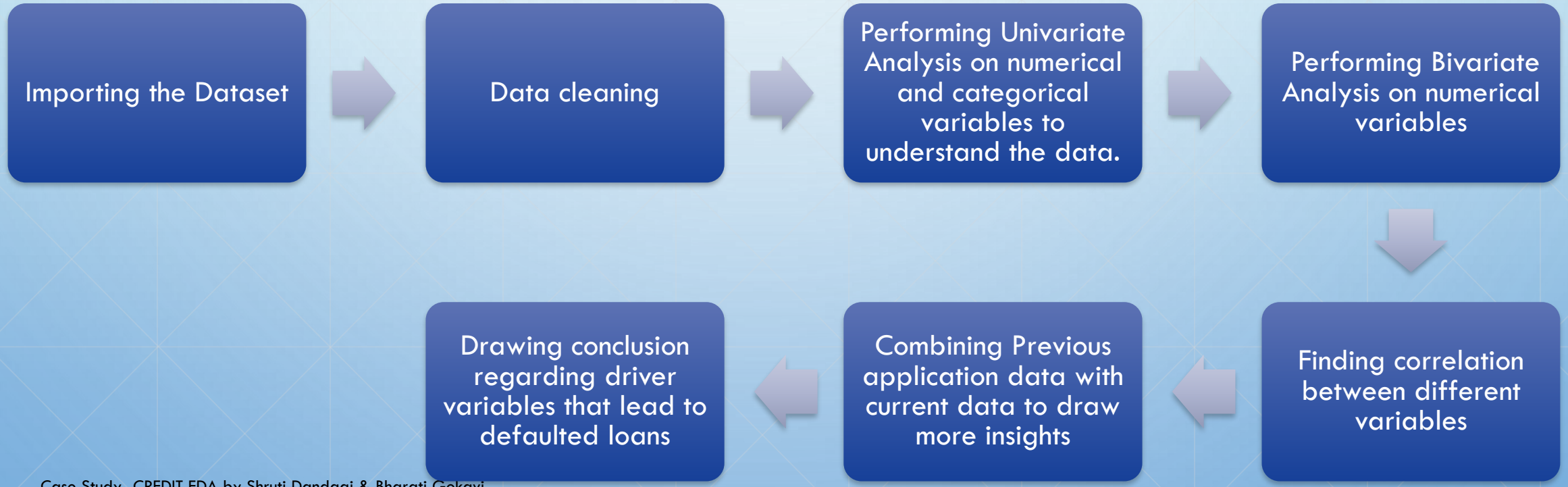
CASE STUDY- CREDIT EDA

BY SHRUTI DANDAGI & BHARATI GOKAVI

OBJECTIVE

- Identify patterns which indicate if a client has difficulty paying their instalments 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.

Steps Taken:



DATA CLEANING

Reading and Inspection

- We read the dataset and inspect the number of Rows, columns, datatypes, Numeric columns.

Check missing values

- Find out the number of Null values in all the columns. Also, find the percentage of Null values in each column. Round off the percentages up to two decimal places.

Impute/Remove missing values

- The null values in some columns are not required and have been dropped
- The columns with higher null values will not help our Analysis. Hence finding them and cleaning the dataset. Here we dropped the missing values which are above 50%

Drop unnecessary columns

- Many of the columns in this data frame are not required for our analysis, So we can drop some columns

Check the Datatypes and Convert if needed

- There are cases where the dataset can present incorrect datatypes for variables.
- For ex: Suppose there is a numeric column and it can display the datatype as object. So we need to check that and correct the datatypes

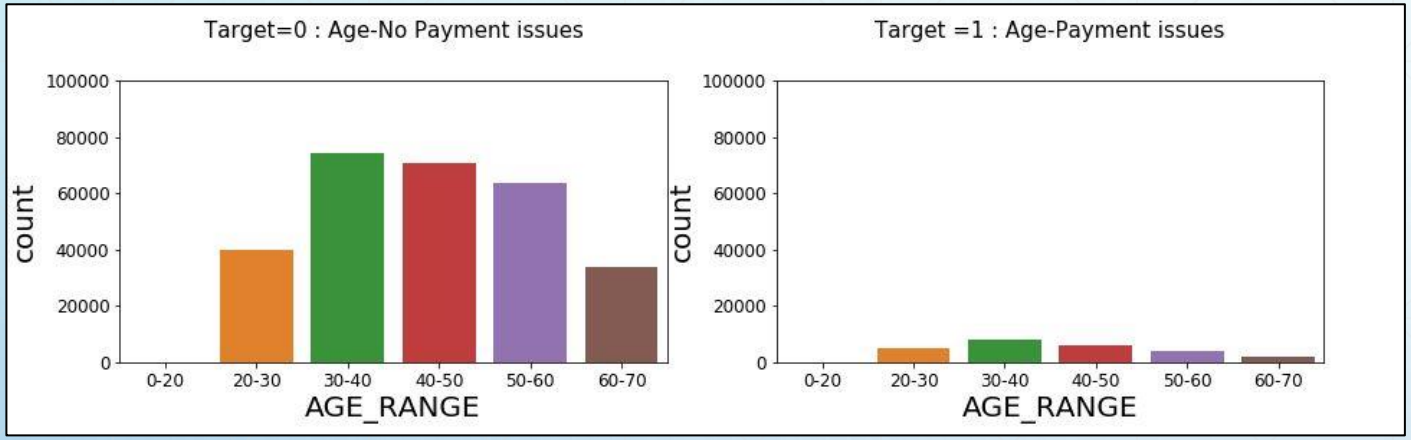
Handling Outliers

- Outliers are values that are much beyond or far from the next nearest data points. These outliers should be treated before investigating data and drawing insights from a dataset.

UNIVARIATE ANALYSIS: FOR TARGET=0, TARGET =1

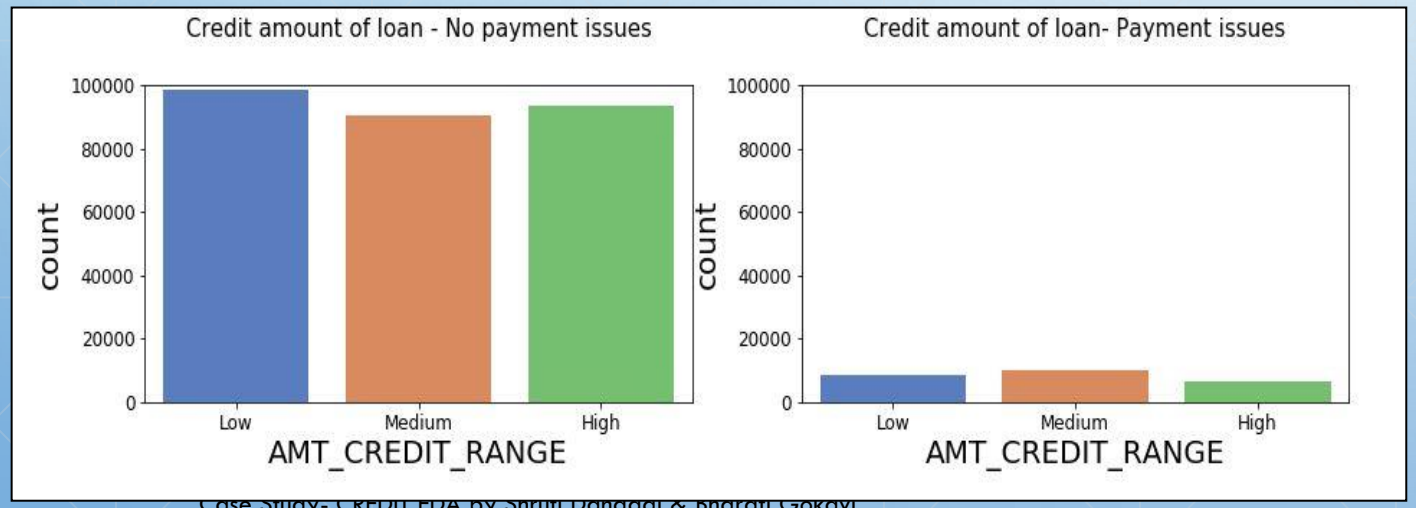
Numeric Variables:

Age



- Here in this plot analysis is done on Age group of the people who are having payment issues and people with no payment issues.
- We can observe that customers belonging to age group 30-40 are able to make payment on time and can be considered while lending loan. Next the people between age 40-60 can be considered.

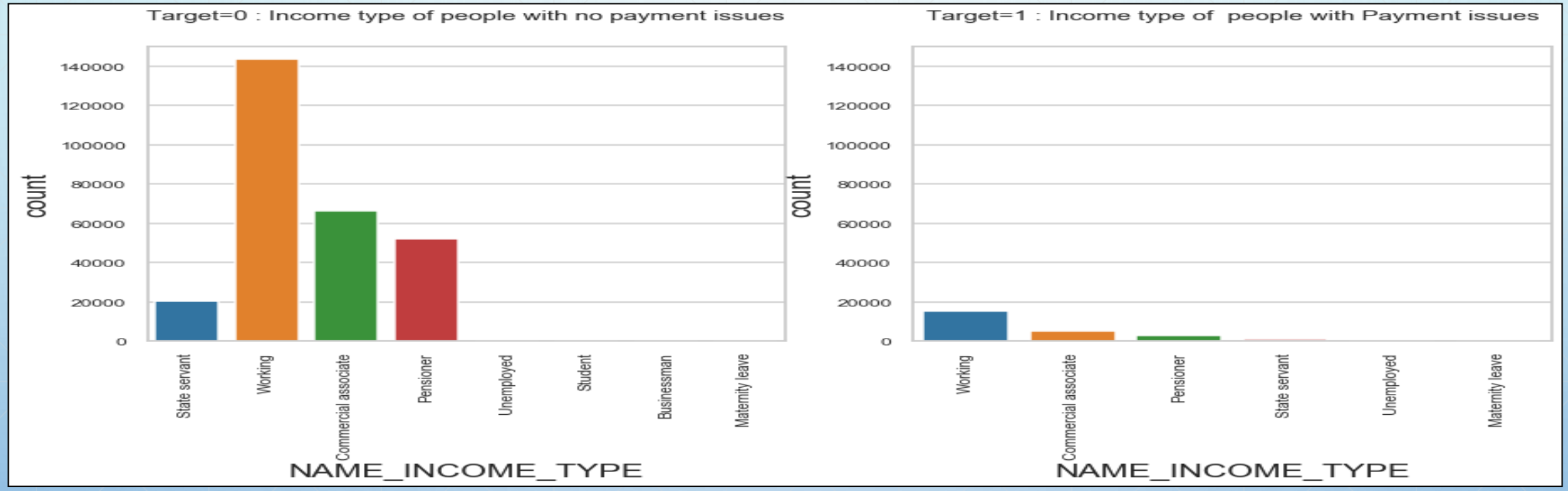
Credit Amount



- Customers with less credit and most likely to make payment.
- Customers with medium credit face payment issues.

CATEGORICAL VARIABLES

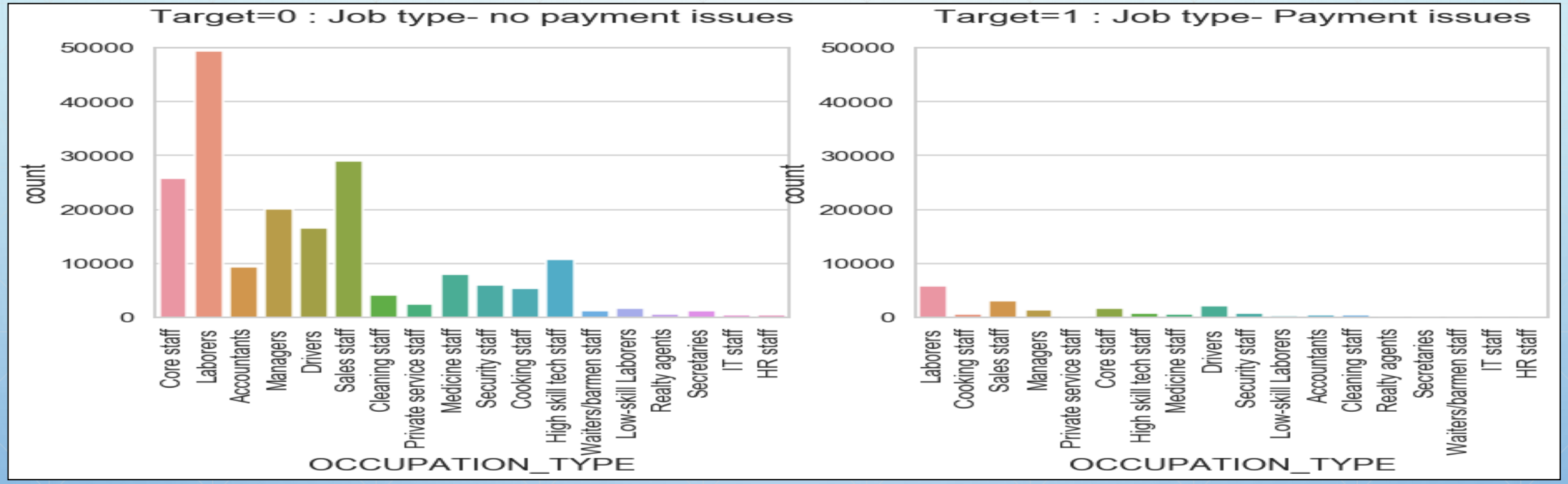
Income Type



- Plot shows the credit amount of loan for clients with payment issues and no payment issues. About 7/8th Working class employees are most likely to make payments on time. Where as 1/8th of working customers face difficulties in making payment.
- On the other side we can also we that comparatively working class employees also face payment issues

CATEGORICAL VARIABLES CONTINUED…

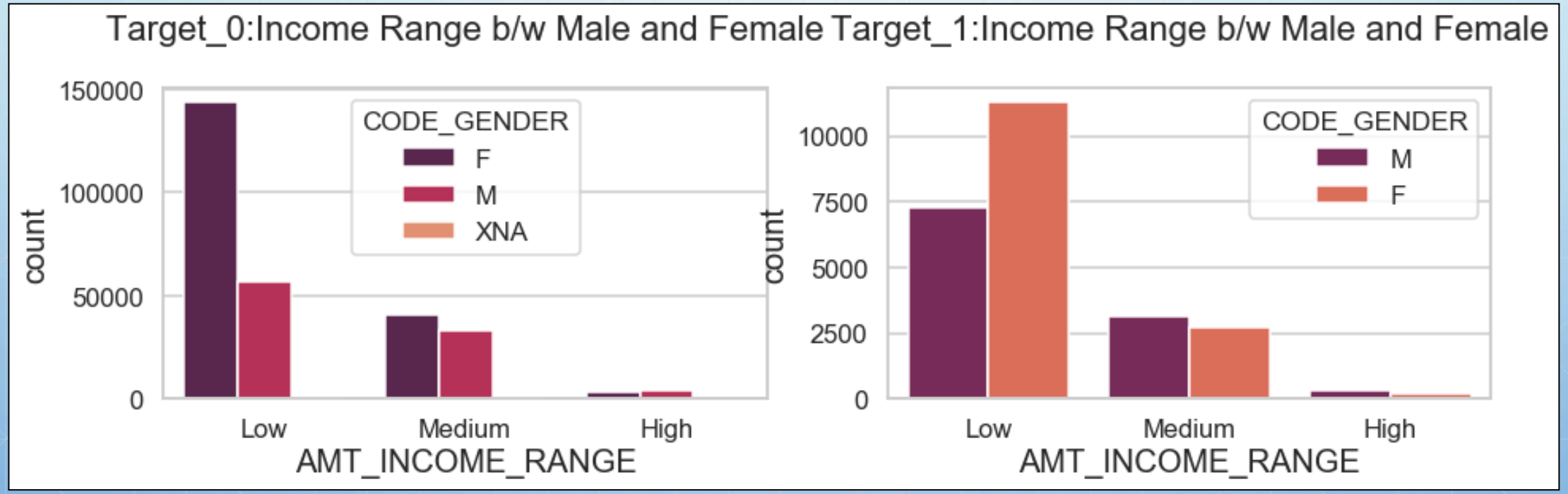
Job Type



- The plot gives the occupation type of customers vs target.
- The plot clearly shows that labourers are most likely to make payment on time whereas HR staff are less likely to make payment on time.

CATEGORICAL VARIABLES CONTINUED…

Income Range

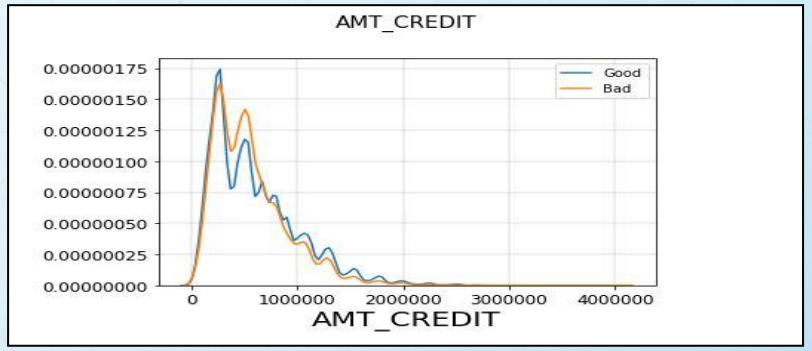


We can see that Females with low income don't have any payment issues

ANALYSE CONTINUOUS COLUMN WITH RESPECT TO THE "TARGET" COLUMN

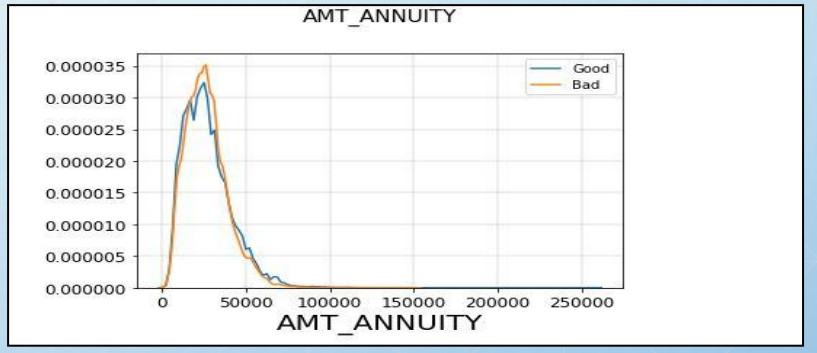
Credit amount

- Credit amount credited doesn't vary much for Target=0 and target =1



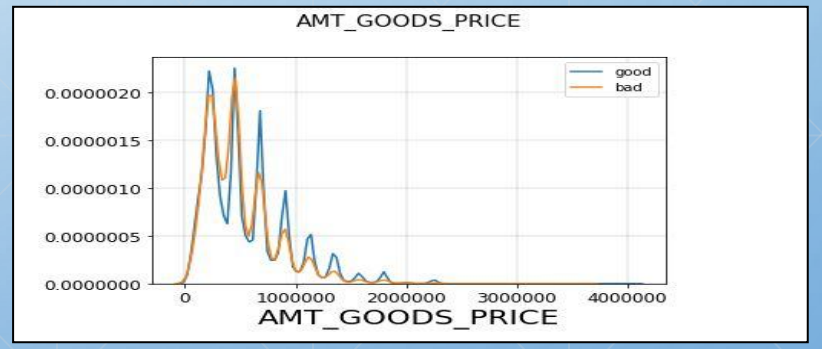
Amount Annuity

- Amount annuity doesn't vary much for target=0 and target=1

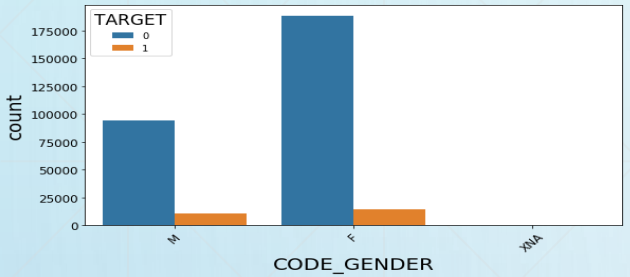


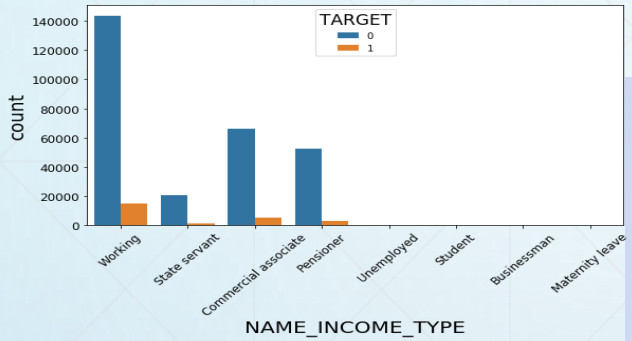
Goods Price

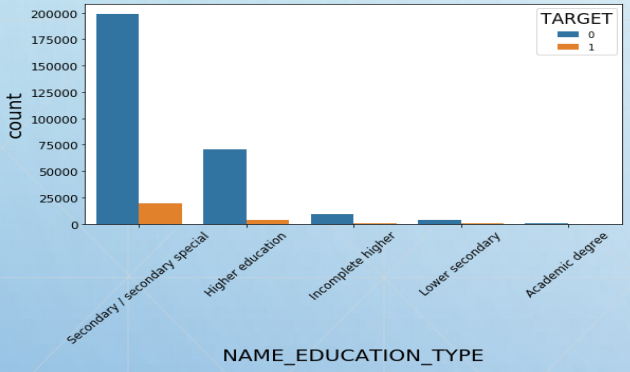
- Goods price based on which the loan is granted doesn't vary much for target 0 and target 1

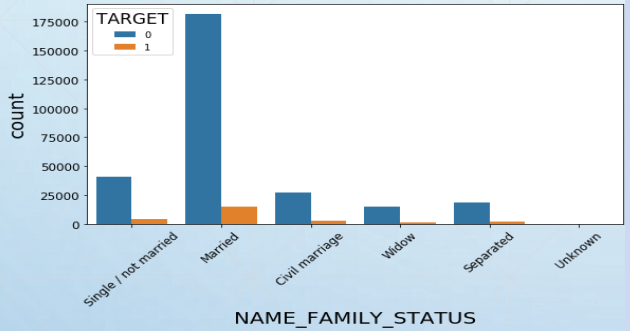


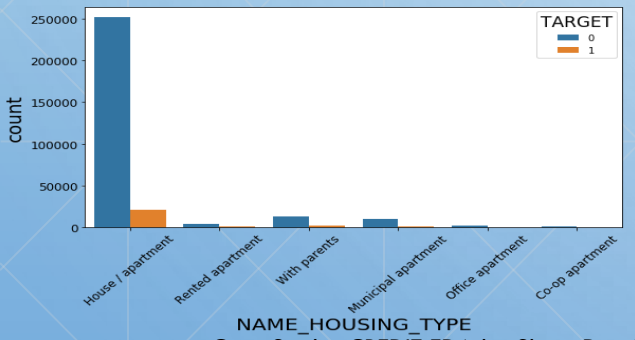
CATEGORICAL COLUMNS WITH RESPECT TO "TARGET" COLUMN

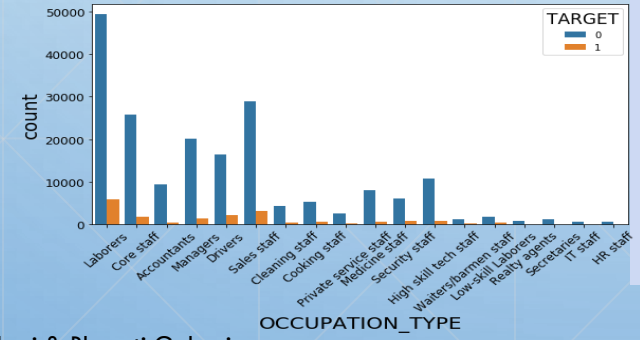












- Female customers pay loan amount on time and banks can target more female customers for lending loan.
- Working customers can be targeted to lend loans as they have higher percentage of making payments on time.
- Customers with secondary education are most likely to make payments when compared to customers with academic degree.
- Married customers have paid loan amount on time when compared to widows.
- Customers owning House/apartment are most likely to make payments on time compared to those living in CO-OP apartment.
- Labourers have high repayment percentage. Hence banks can think of lending small amount loans to them.

CORRELATION MATRIX FOR TARGET_0 AND TARGET_1 DATA FRAME

Target=1

	VAR1	VAR2	Correlation
398	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
148	AMT_GOODS_PRICE	AMT_CREDIT	0.98
423	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87
124	AMT_ANNUITY	AMT_CREDIT	0.75
149	AMT_GOODS_PRICE	AMT_ANNUITY	0.75
224	DAYS_EMPLOYED	DAYS_BIRTH	0.58
399	OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.34
374	DEF_30_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.33
248	DAYS_REGISTRATION	DAYS_BIRTH	0.29
422	DEF_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.26

Target=0

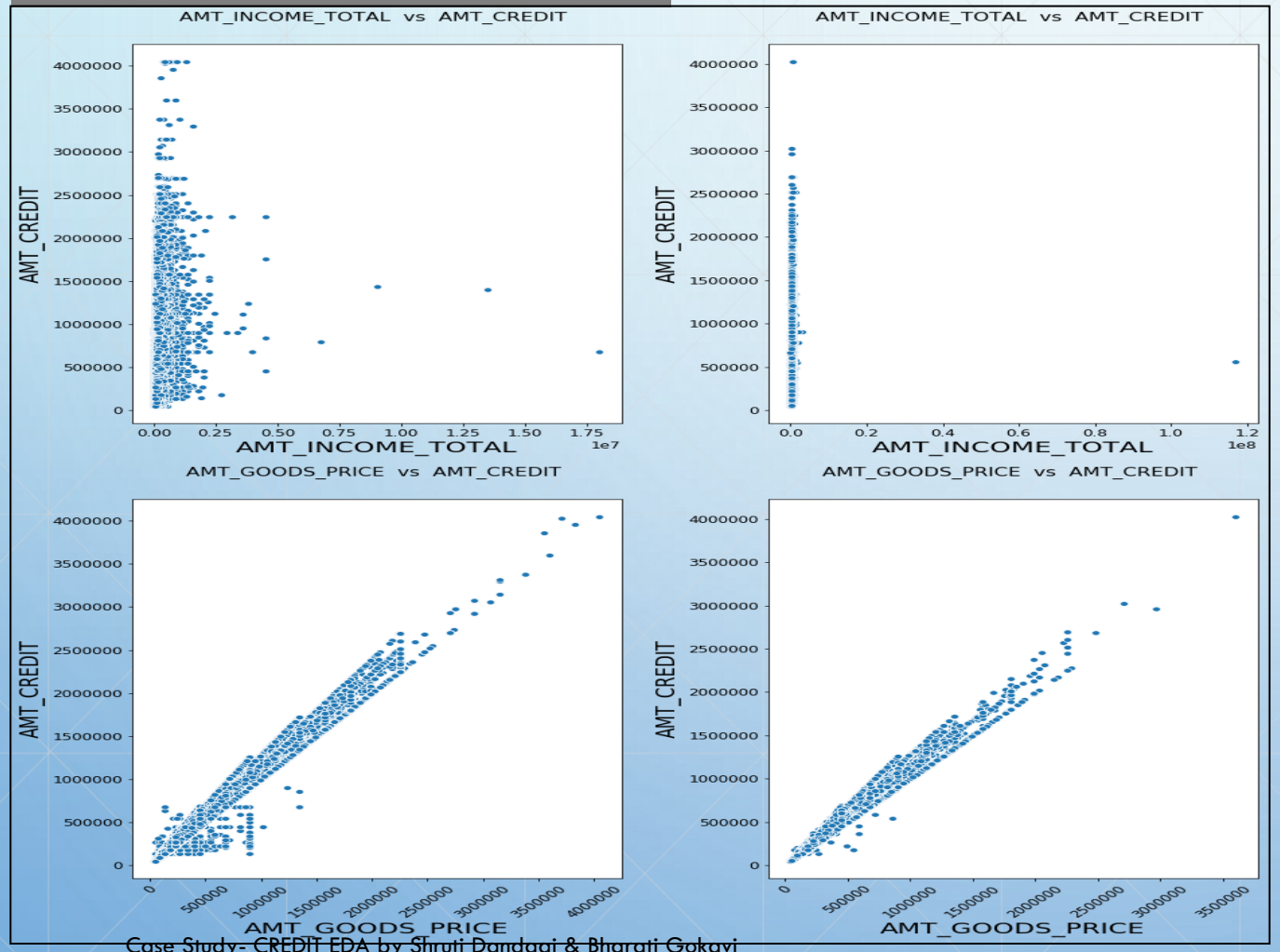
	VAR1	VAR2	Correlation
398	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
148	AMT_GOODS_PRICE	AMT_CREDIT	0.99
423	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86
149	AMT_GOODS_PRICE	AMT_ANNUITY	0.78
124	AMT_ANNUITY	AMT_CREDIT	0.77
224	DAYS_EMPLOYED	DAYS_BIRTH	0.63
123	AMT_ANNUITY	AMT_INCOME_TOTAL	0.42
147	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.35
194	DAYS_BIRTH	CNT_CHILDREN	0.34
99	AMT_CREDIT	AMT_INCOME_TOTAL	0.34

- We can see that for Target_0 and Target_ 1 data frame, Social circle for 30 days and 60 days are most co-related and Goods price and Loan amount credit are highly correlated. Then we have Goods price and amount annuity on 4th place

- So the variables correlated in target_0 data frame and target_1 data frame are same with slightly varying correlation values¶

BIVARIATE ANALYSIS FOR TARGET=0 AND TARGET 1

Numeric- Numeric Analysis

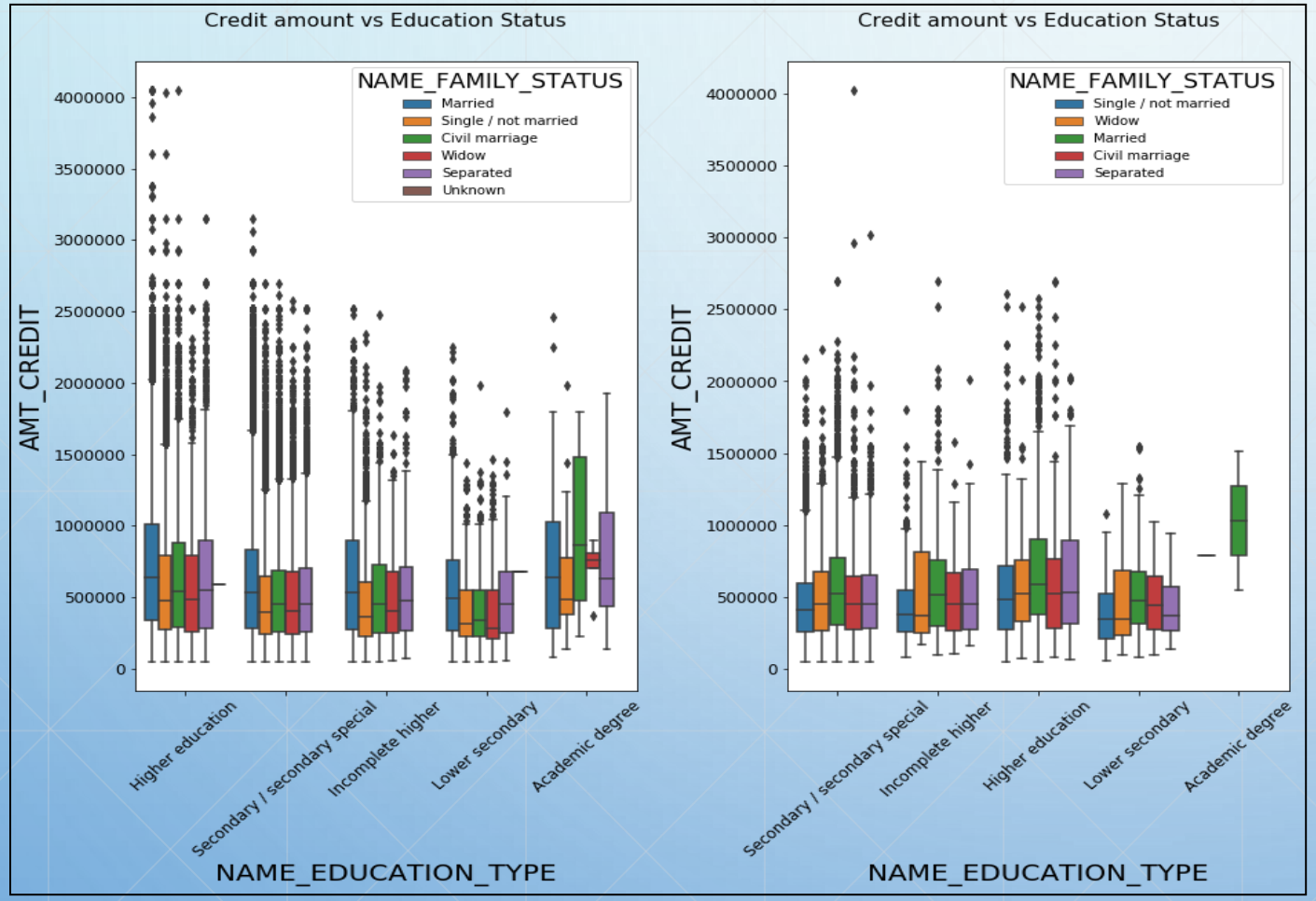


- THOSE WHO HAVE PAID THE LOAN AMOUNT ON/WITHIN TIME ARE MORE LIKELY TO GET HIGHER CREDITS THAN THOSE WHO DIDN'T PAY/DID LATE PAYMENTS.

- PEOPLE WHO HAVE HIGHER GOODS PRICE AND HAVE MADE PAYMENTS ON TIME HAVE HIGHER CREDITS THAN THOSE WITH HIGHER GOODS PRICE BUT DIDN'T PAY LOAN.

BIVARIATE ANALYSIS CONTINUED…

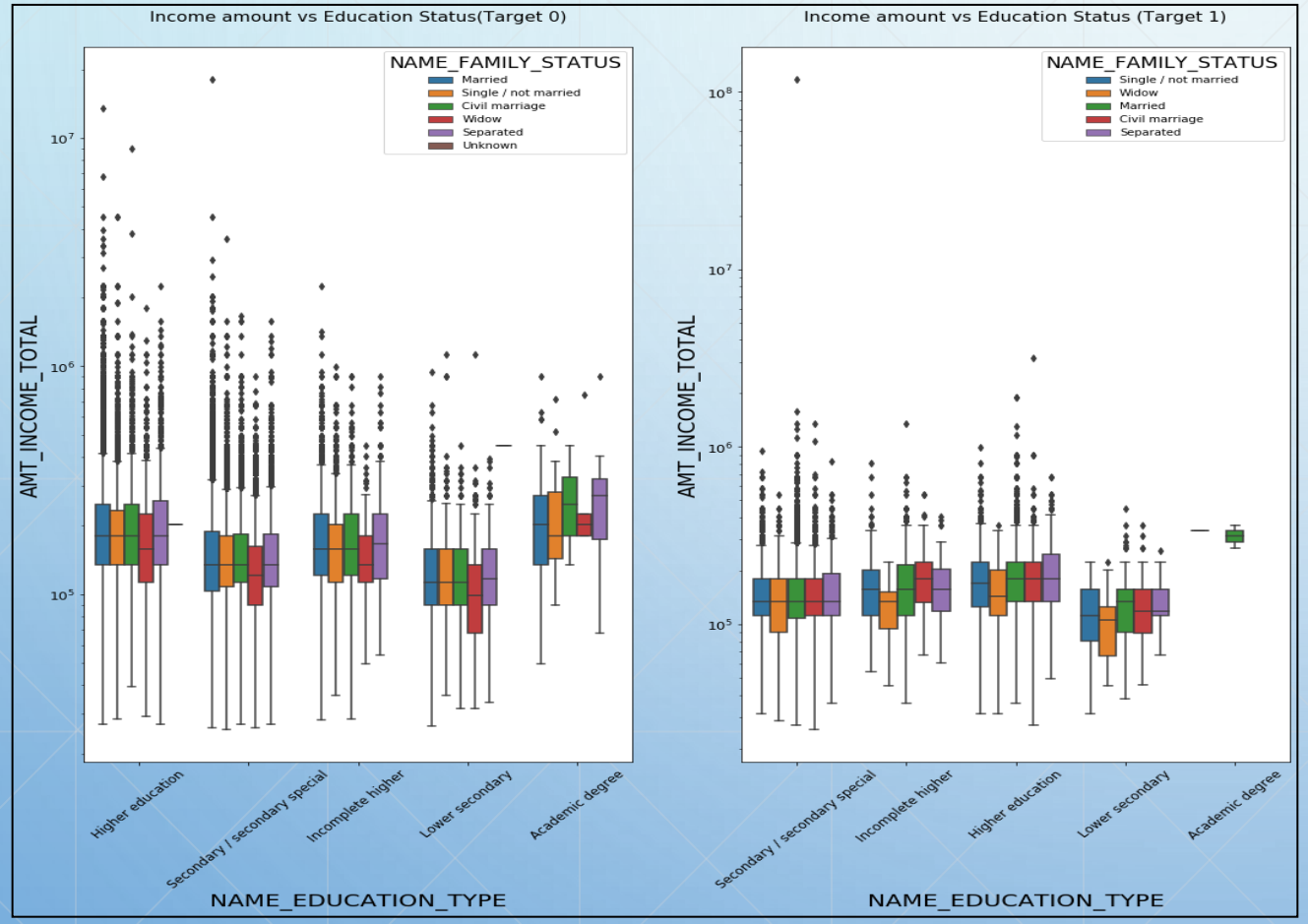
Numerical categorical analysis



- Some of the highly educated, married people are having credits higher than those who have done lower secondary education.
- Those with higher education have higher credits and are more likely to make payments on time.
- More number of outliers are seen in higher education.4.The people with secondary and secondary special education are less likely to make payments on time.

NUMERICAL-CATEGORICAL ANALYSIS CONTINUED...

Numerical categorical analysis



- From the above plots,
- We can see that Higher education has many outliers.
- People with higher education have higher income and don't have difficulties in making loan payment.
- People with higher education whoever lesser income are unable to pay the loan.
- Hence we can conclude that, people with Higher income are most likely to make payments.

Merge the Present application data with Previous application data

Merge the application data with the previous application

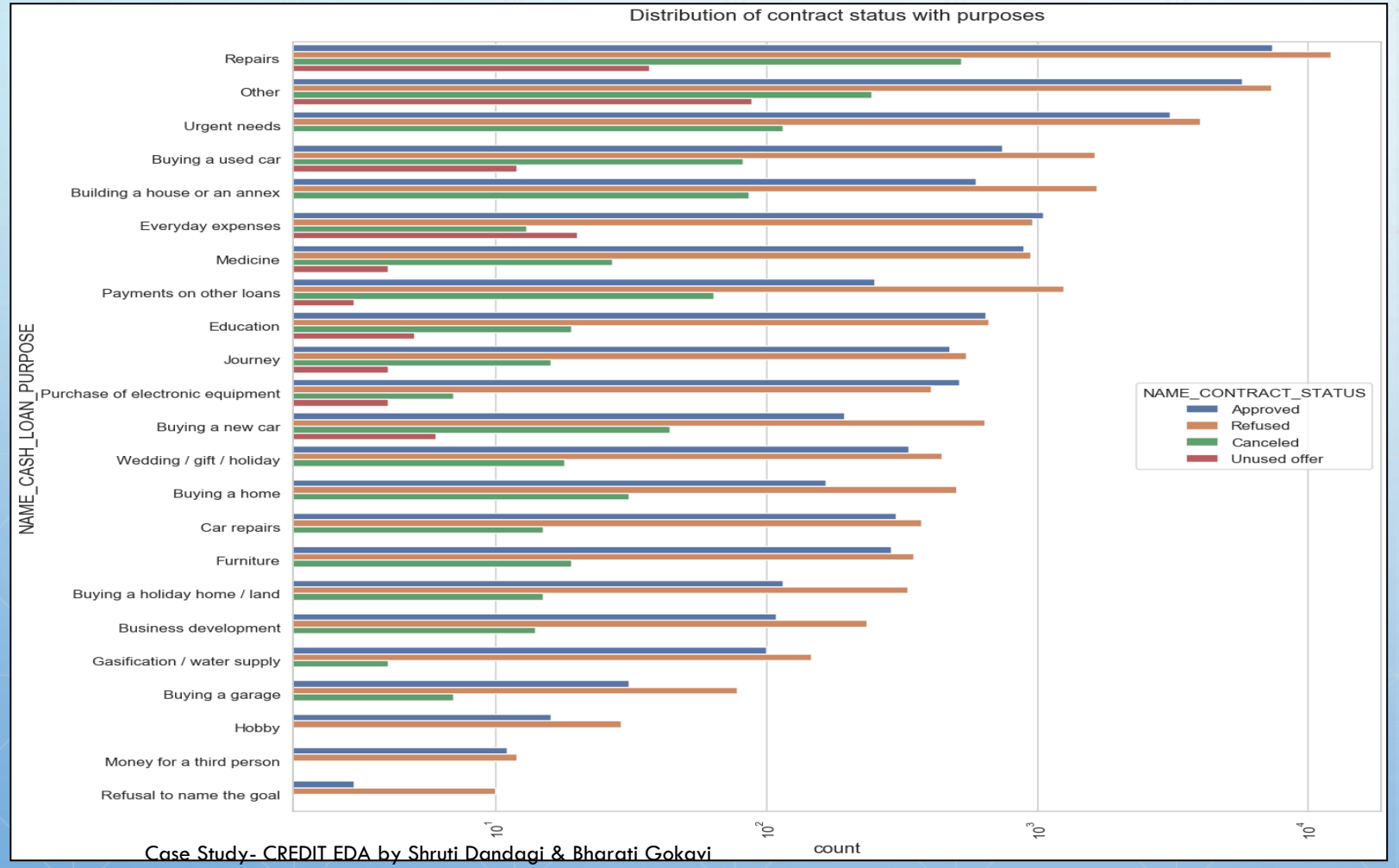
Drop the unnecessary columns

Perform Univariate and Bivariate Analysis

Conclusion

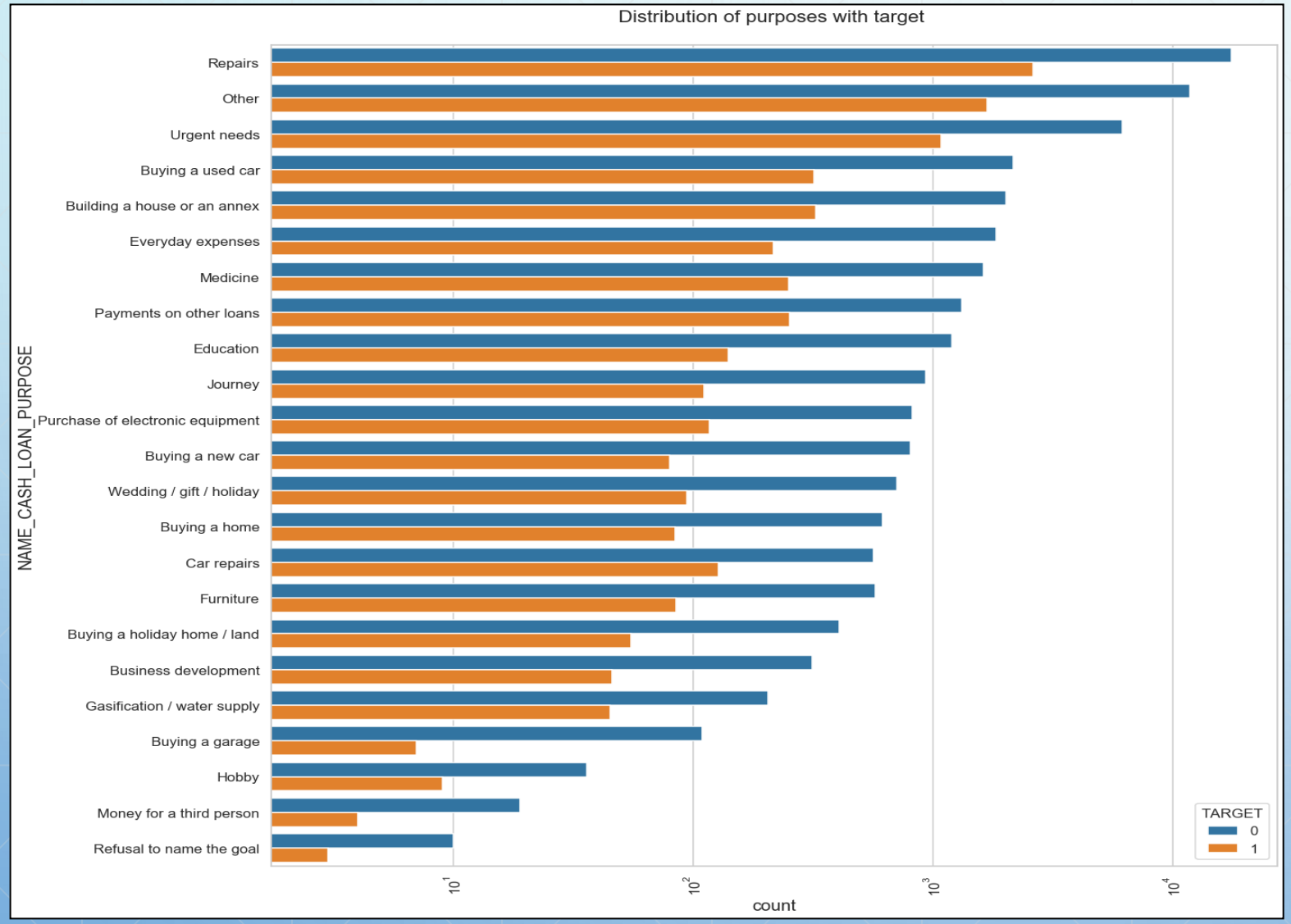
UNIVARIATE ANALYSIS AFTER MERGING THE TWO DATASETS

Distribution of Contract status with purposes



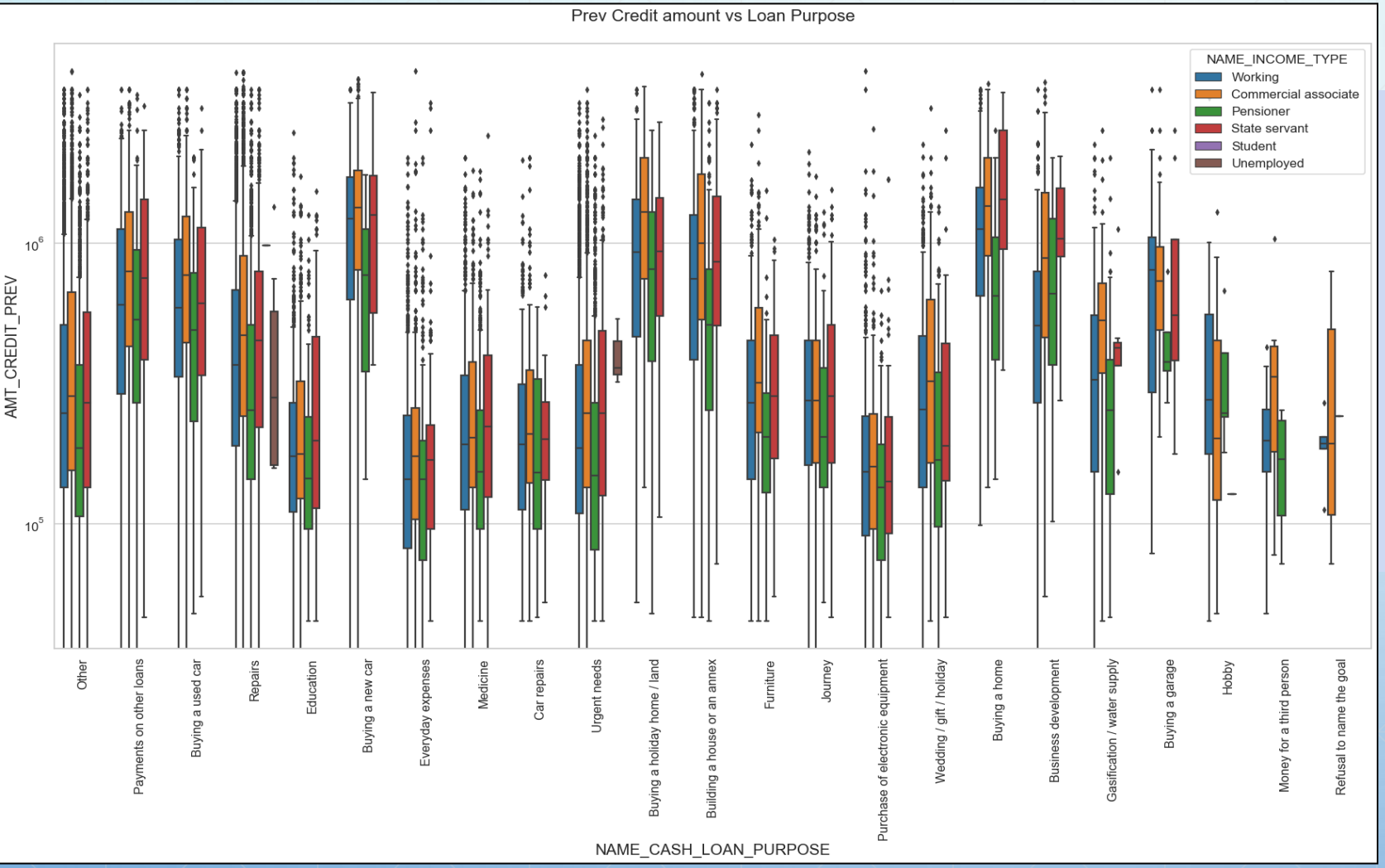
- Most rejection of loans came from purpose 'Repairs'.
- For education purposes we have equal number of approves and rejection Paying other loans and buying a new car is having significant higher rejection than approves.

Distribution of Purposes with Target



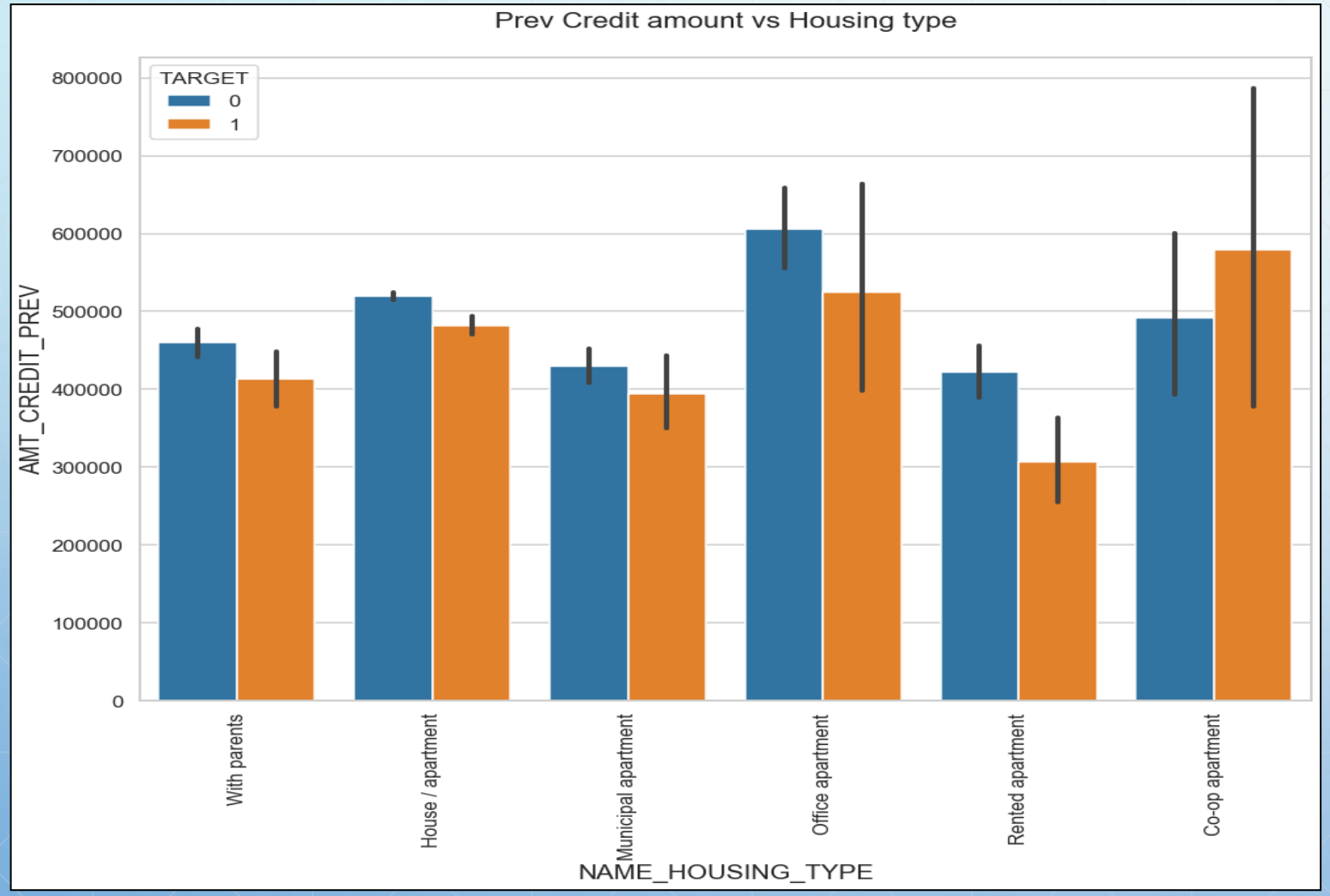
- Few points we can conclude from above plot:
- Loan purposes with 'Repairs' are facing more difficulties in payment on time.
- There are few places where loan payment is significantly higher than facing difficulties.
- They are 'Buying a garage', 'Business development, 'Buying land', 'Buying a new car' and 'Education' Hence we can focus on these purposes for which the client is having for minimal payment difficulties

BIVARIATE ANALYSIS



- The credit amount of Loan purposes like 'Buying a home, Buying a land, Buying a new car and 'Building a house' is higher.
- Income type of state servants have a significant amount of credit applied
- Money for third person or a Hobby is having less credits applied for.

Bivariate continued..



- Here for Housing type, office apartment is having higher credit of target 0 and co-op apartment is having higher credit of target 1.
- So, we can conclude that bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment. Bank can focus mostly on housing type with parents or House\apartment or municipal apartment for successful payments.

CONCLUSION

- Banks should focus more on contract type ‘student’ ,’pensioner’ and ‘businessman’ with housing ‘type other than ‘co-op apartment’ for successful payments.
- And on age group 30-40 as they are also likely to make successful payments on time.
- Banks should focus less on income type ‘working’ as they are having most number of unsuccessful payments.
- Also with loan purpose ‘repair’ is having higher number of unsuccessful payments on time .
- Get as much as clients from housing type ‘with parents’ as they are having least number of unsuccessful payments.

THANK YOU