Presentation on Credit EDA Case Study

D. MARTIN ROBERT DURAI

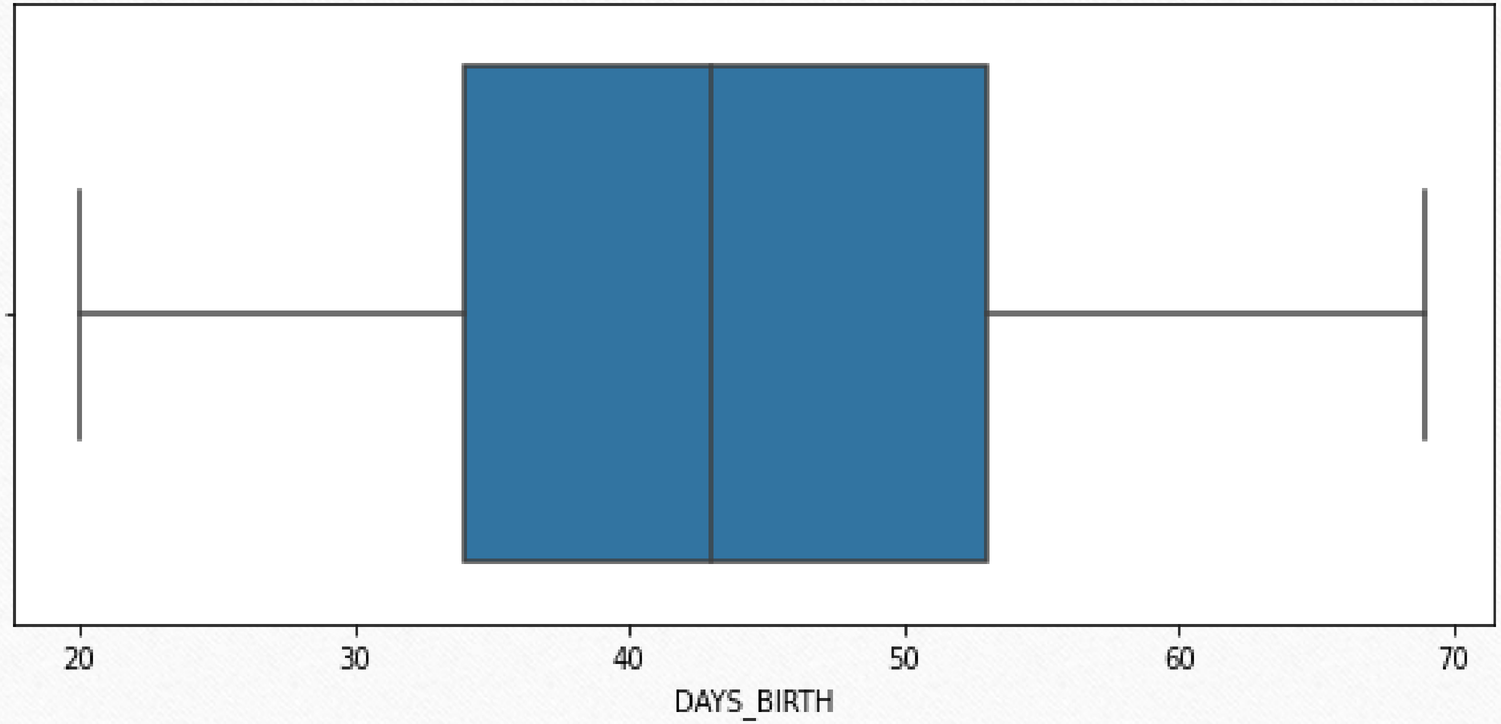
Purpose

- Credit risk analysis will help the company to make a decision for loan approval based on the applicant’s profile.
- When controls loss of business to the company and avoid financial loss for the company.

Steps

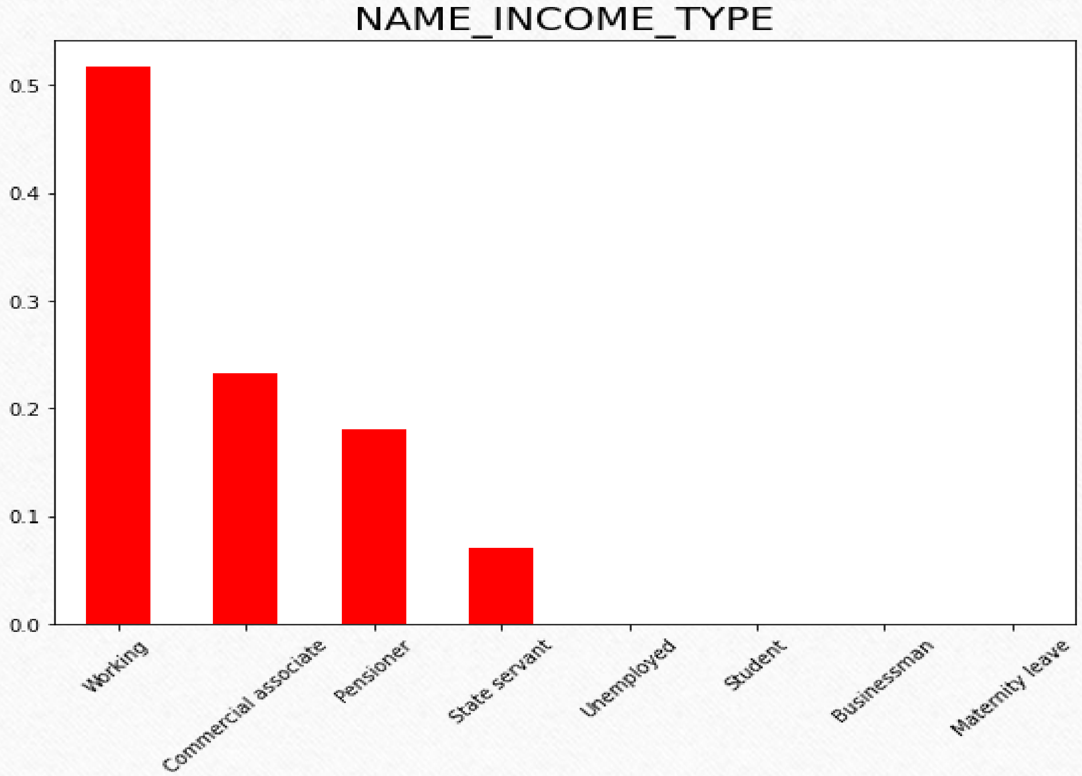
- Data understanding and sourcing.
- Checking for Data quality and Binning.
- Checking for errors in data types, Univariate analysis& Bivariate analysis, correlation.
- Merging of application data with previous application data.
- Data analysis by univariate, Bivariate analysis and correlation.
- Insights

Analyze DAYS_BIRTH column for outliers in Application data



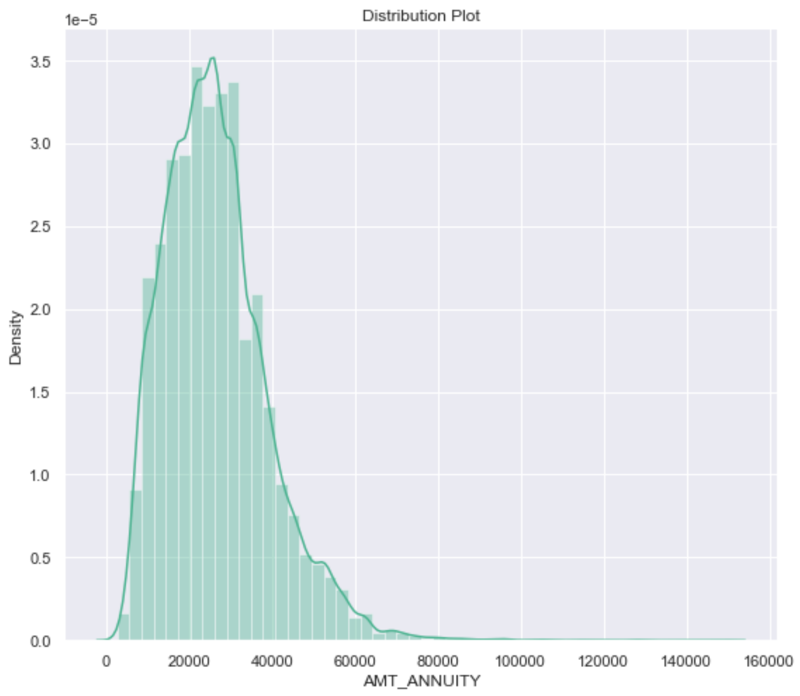
Here the values are stored as number of days since birth which is not easily comprehensible, so we can add a new column age in years by transforming this and just analyze that.

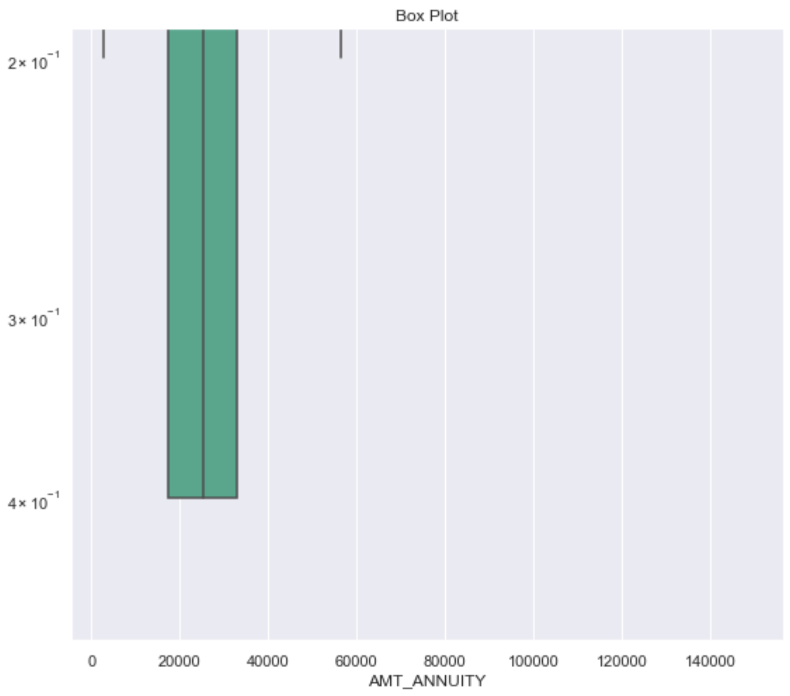
Univariate Analysis of Categorical Variables in Application data



We observe a decrease in the percentage of Payment Difficulties who are pensioner and an increase in the percentage of Payment Difficulties who are working when compared the percentages of both Payment Difficulties and non-Payment Difficulties.

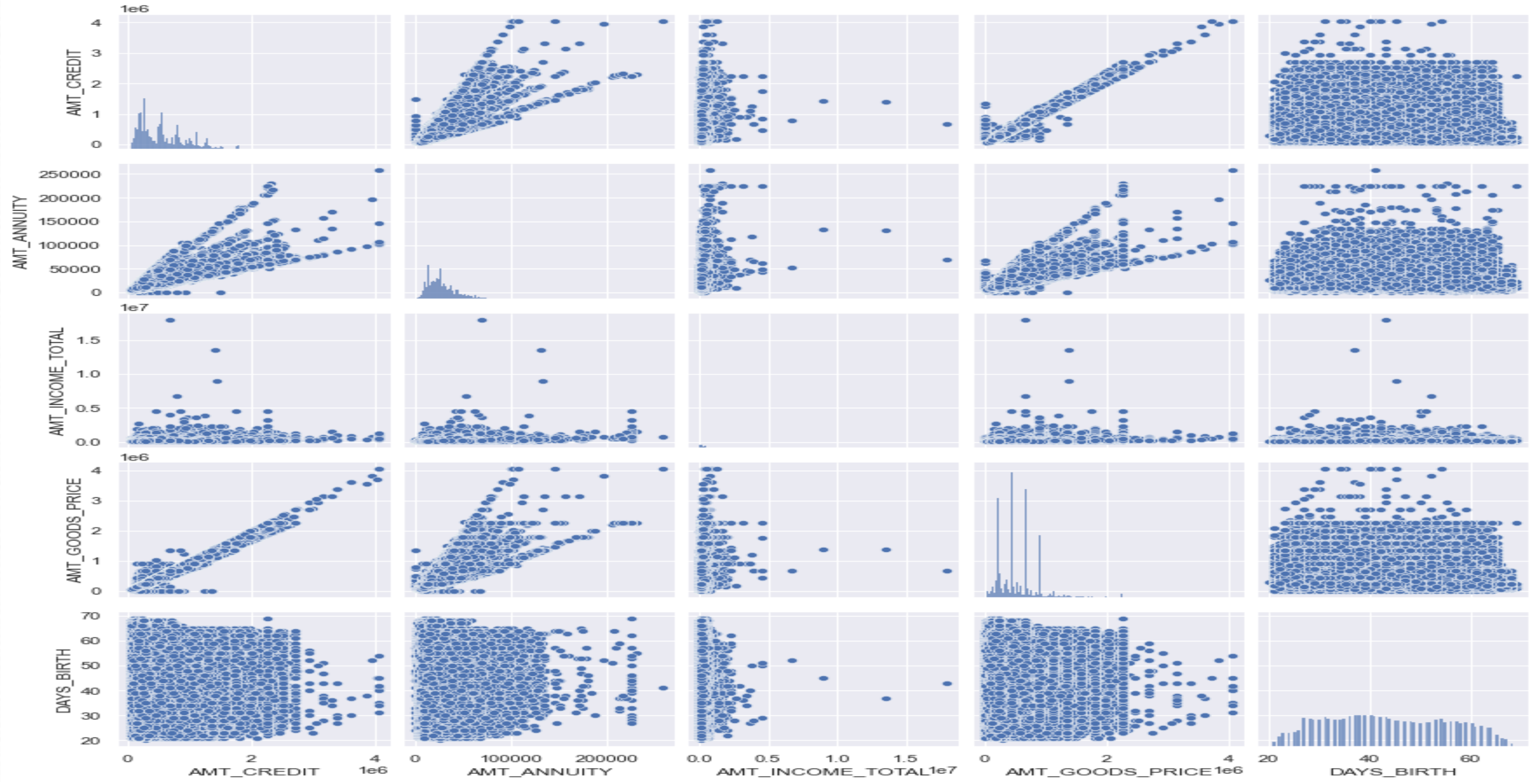
Univariate Analysis of Numerical Variables in Application data





We can observe some outliers and the first quartile is bigger than third quartile for annuity amount which means most of the annuity clients are from first quartile.

Bivariate Analysis of Numerical vs Numerical Variables in Application data



Correlation for Loan- Non Payment Difficulties in Application data

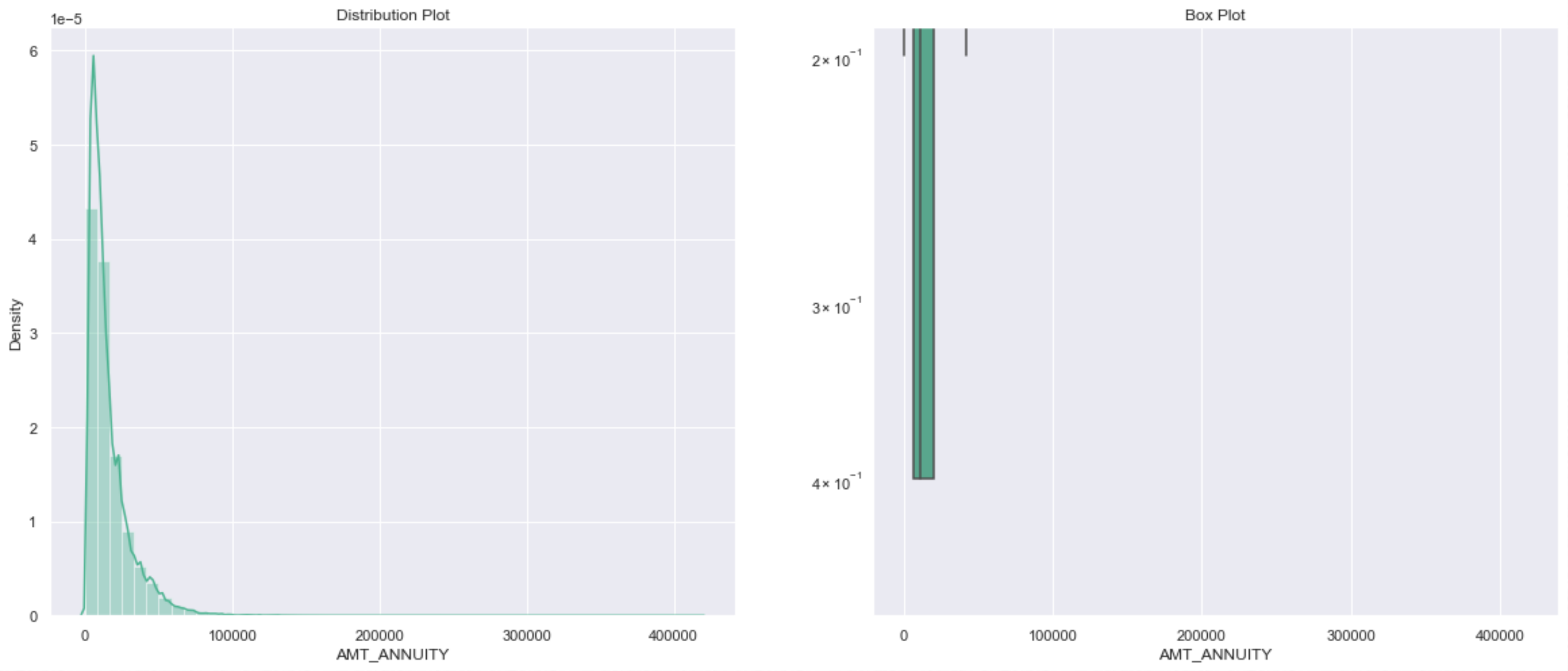
	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	AMT_ANNUITY	DAYS_EMPLOYED	DAYS_BIRTH	DAYS_REGISTRATION	DAYS_ID_PUBLISH	AMT_CREDIT
AMT_GOODS_PRICE	1.000000	0.349462	0.776686	-0.068609	0.044552	-0.015916	0.003649	0.987250
AMT_INCOME_TOTAL	0.349462	1.000000	0.418953	-0.140392	-0.062494	-0.064937	-0.022896	0.342799
AMT_ANNUITY	0.776686	0.418953	1.000000	-0.104978	-0.012254	-0.039436	-0.014113	0.771309
DAYS_EMPLOYED	-0.068609	-0.140392	-0.104978	1.000000	0.626028	0.214511	0.276663	-0.070104
DAYS_BIRTH	0.044552	-0.062494	-0.012254	0.626028	1.000000	0.333025	0.270804	0.047366
DAYS_REGISTRATION	-0.015916	-0.064937	-0.039436	0.214511	0.333025	1.000000	0.100236	-0.013477
DAYS_ID_PUBLISH	0.003649	-0.022896	-0.014113	0.276663	0.270804	0.100236	1.000000	0.001464
AMT_CREDIT	0.987250	0.342799	0.771309	-0.070104	0.047366	-0.013477	0.001464	1.000000

We observe that there is a high correlation between credit amount and goods price. There appears to be some deviancies in the correlation of Loan-Payment Difficulties and Loan- Non Payment Difficulties such as credit amount v/s income.

Top 10 Correlation for client with payment difficulties in Application data

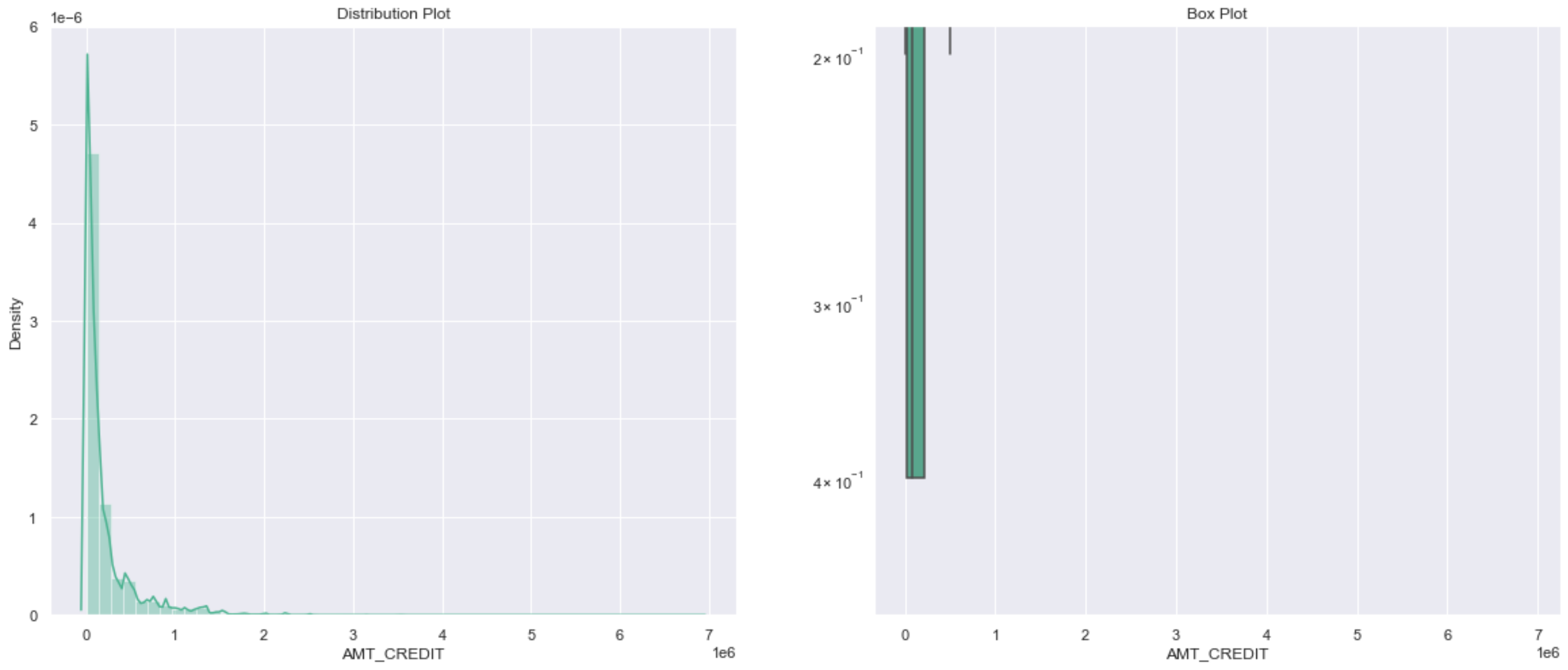
	VAR1	VAR2	CORRELATION	CORR_ABS
56	AMT_CREDIT	AMT_GOODS_PRICE	0.983103	0.983103
16	AMT_ANNUITY	AMT_GOODS_PRICE	0.752699	0.752699
58	AMT_CREDIT	AMT_ANNUITY	0.752195	0.752195
35	DAYS_BIRTH	DAYS_EMPLOYED	0.582438	0.582438
17	AMT_ANNUITY	AMT_INCOME_TOTAL	0.398260	0.398260
8	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.327484	0.327484
57	AMT_CREDIT	AMT_INCOME_TOTAL	0.325386	0.325386
44	DAYS_REGISTRATION	DAYS_BIRTH	0.289135	0.289135
52	DAYS_ID_PUBLISH	DAYS_BIRTH	0.252272	0.252272
51	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.229102	0.229102

Univariate analysis of numerical columns in AMT_ANNUITY of Previous Application data



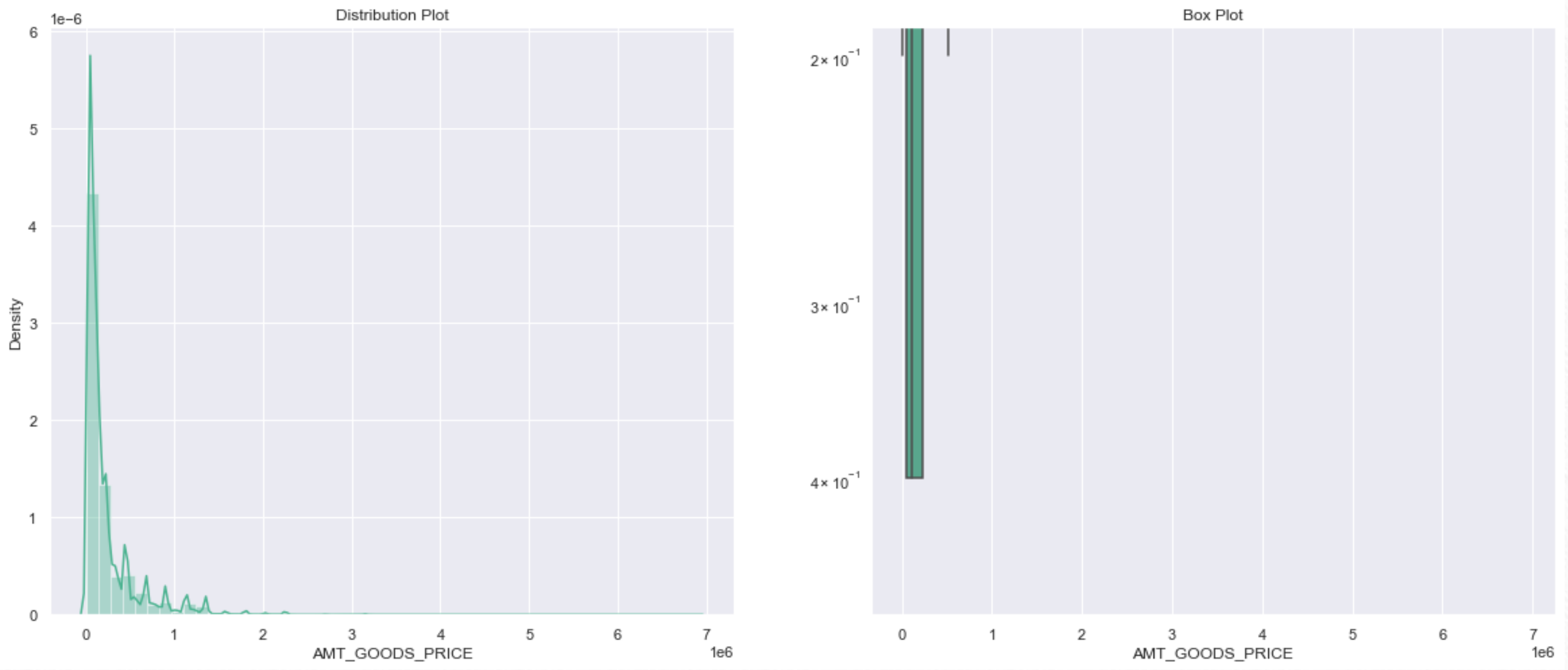
We observe that there are some outliers and the curve is not normal or a bell curve.

Univariate analysis of numerical columns in AMT_CREDIT of Previous Application data



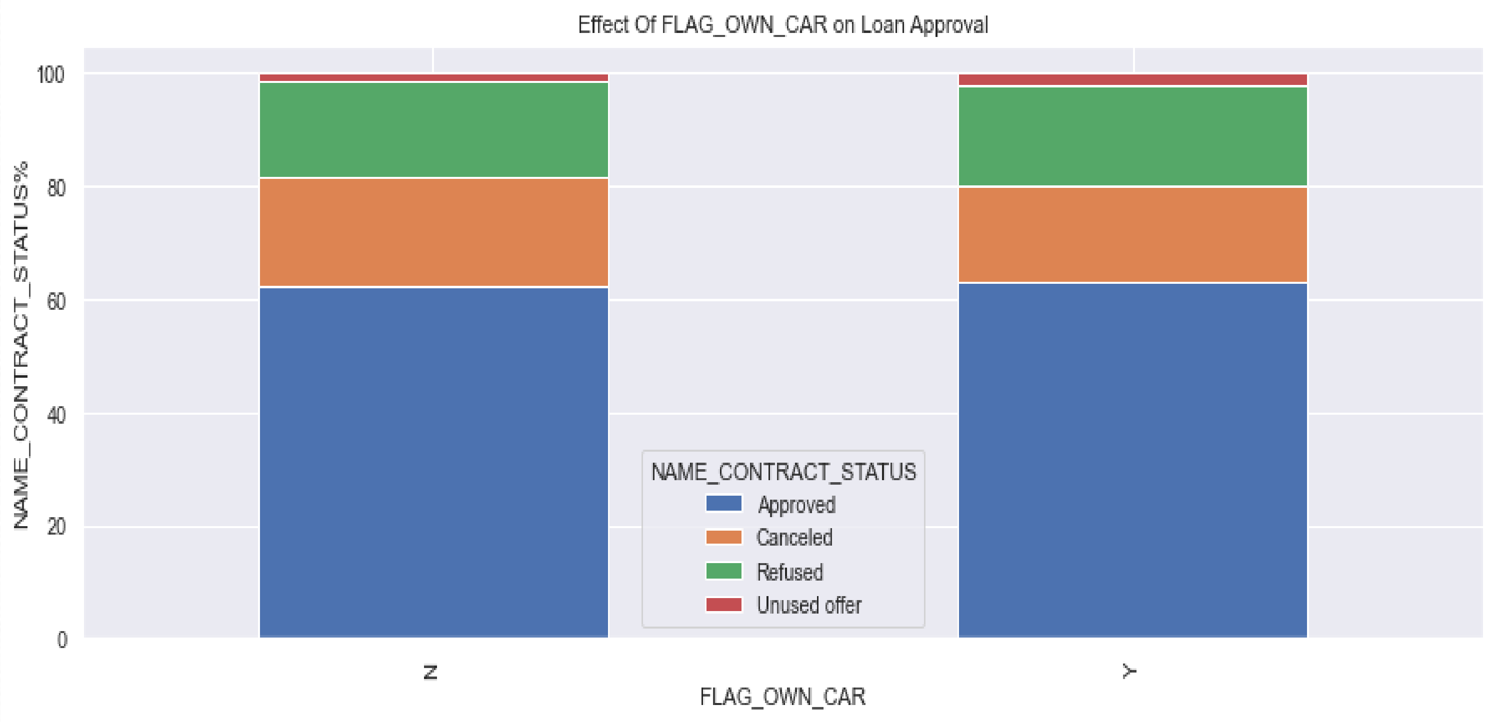
We observe that there are some outliers and the curve is not normal or a bell curve.

Univariate analysis of numerical columns in AMT_GOODS_PRICE of Previous Application data



We observe that there are some outliers and the curve .is not normal or a bell curve.

Univariate Merged on FLAG_OWN_CAR in Application data

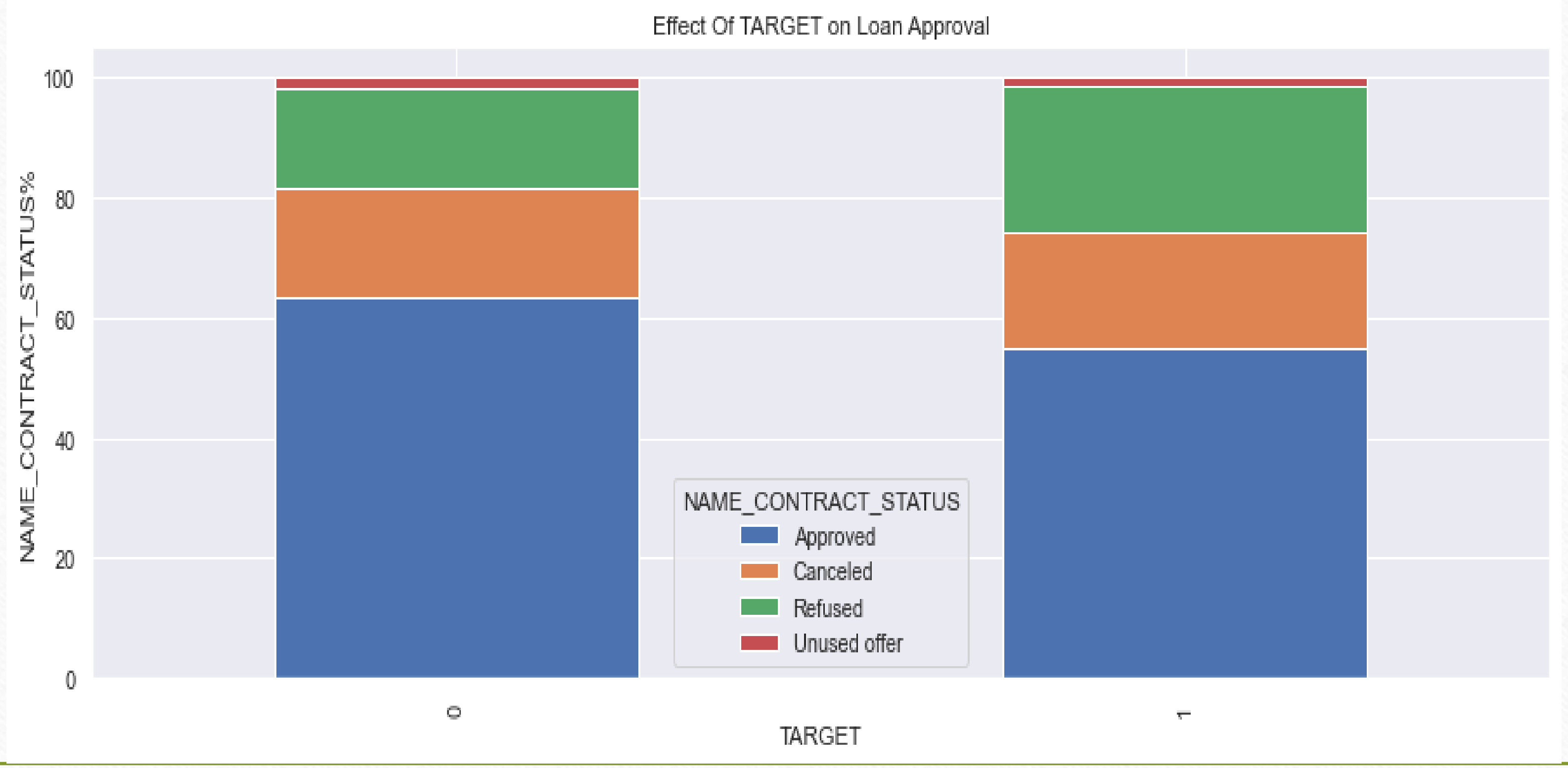


We see that car ownership doesn't have any effect on application approval or rejection. But we saw earlier that the people who has a car has lesser chances of default. The bank can add more weightage to car ownership while approving a loan amount.

Univariate Merged on CODE_GENDER in Application data

We see that code gender doesn't have any effect on application approval or rejection.
But we saw earlier that female have lesser chances of default compared to males. The bank can add more weightage to female while approving a loan amount.

Univariate Merged on TARGET in Application data



We can see that the people who were approved for a loan earlier, defaulted less often where as people who were refused a loan earlier have higher chances of defaulting.

Insights for Application data

- The count of ‘Maternity Leave’ in ‘NAME_INCOME_TYPE’ is very less and it also has maximum % of payment difficulties- around 40%. Hence, client with income type as ‘Maternity leave’ are the driving factors for Loan Defaulters.
- The count of ‘Low skilled Laborers’ in ‘OCCUPATION_TYPE’ is comparatively very less and it also has maximum % of payment difficulties- around 17%. Hence, client with occupation type as ‘Low skilled Laborers’ are the driving factors for Loan Defaulters.
- The count of ‘Lower Secondary’ in ‘NAME_EDUCATION_TYPE’ is comparatively very less and it also has maximum % of payment difficulties- around 11%. Hence, client with education type as ‘Lower Secondary’ are the driving factors for Loan Defaulters.

Insights for Previous Application data

- The count of ‘Refusal to name the goal’ in ‘NAME_CASH_LOAN_PURPOSE’ is comparatively very less and it also has maximum % of payment difficulties- around 23%. Hence, clients who have ‘Refused to name the goal’ for cash loan in previous application are the driving factors for Loan Defaulters.
- The count of ‘Refused’ in ‘NAME_CONTRACT_STATUS’ is comparatively less and it also has maximum % of payment difficulties- around 12%. Hence, client with contract status as ‘Refused’ in previous application are the driving factors for Loan Defaulters.
- The count of ‘Revolving Loans’ in ‘NAME_CONTRACT_TYPE’ is comparatively very less and it also has maximum % of payment difficulties- around 10%. Hence, client with contract type as ‘Revolving loans’ in previous application are the driving factors for Loan Defaulters.
- It can be observed from the graph that Clients with 'Revolving loans' and with 'Refused' previous application tend to have more % of payment difficulties in current application. Since the count of both 'Revolving loans' and ‘Refused’ is comparatively less(from the graphs in previous slide), clients with ‘Revolving Loans’ and ‘Refused’ previous application are driving factors for Loan Defaulters


Thank you!