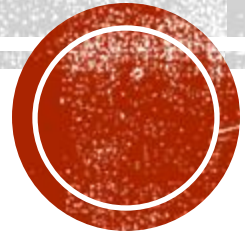
EDA CASE STUDY

CREDIT RISK ANALYSIS



Presented By -

Rahul Chopra

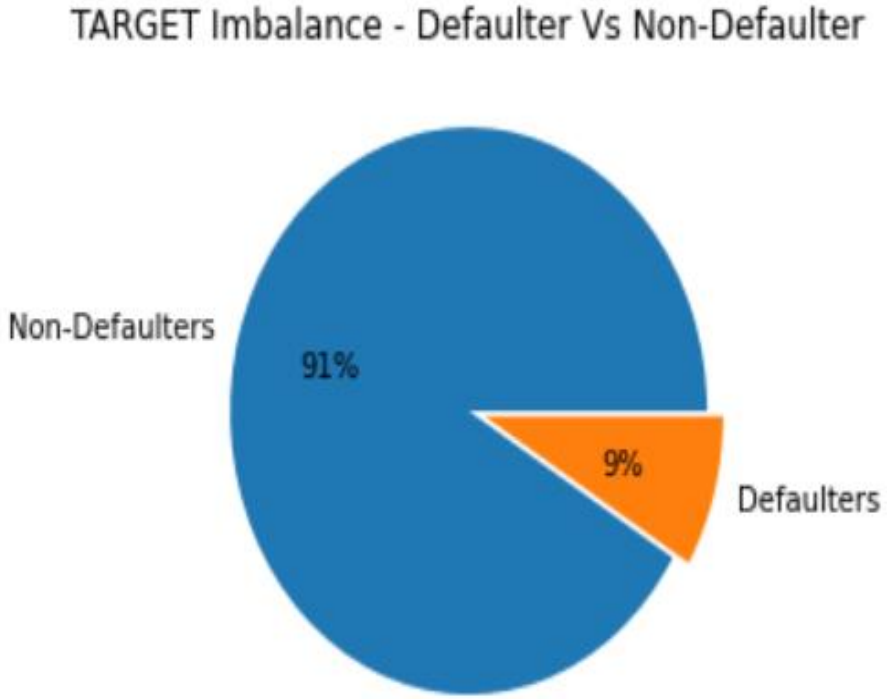
Sonal Hedaoo

PURPOSE OF THE CREDIT CASE STUDY

- In Credit risk case study, we develop a basic understanding of risk analytics in banking and financial services sectors and understand how these industries make decisions and how the data is used to minimize the risk of losing money while lending to customers which controls loss of business.

OPERATIONS PERFORMED

- Data Understanding and Sourcing.
- Check Missing Values in both New and Previous Data sets.
- Check for Outliers and Bining.
- Check for data imbalance and did univariate and Bivariate analysis.
- Combined both the Data Stets (New application data with previous application data).
- Analysed Merged Data Set by univariate and Bivariate analysis.
- Insights.

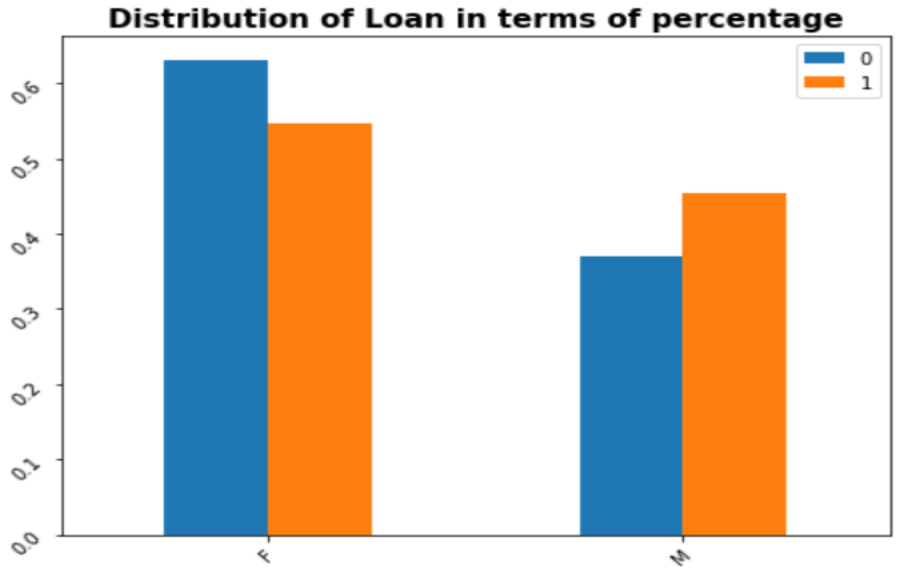


PROPORTION OF DEFAULTERS VS NON-DEFAULTERS

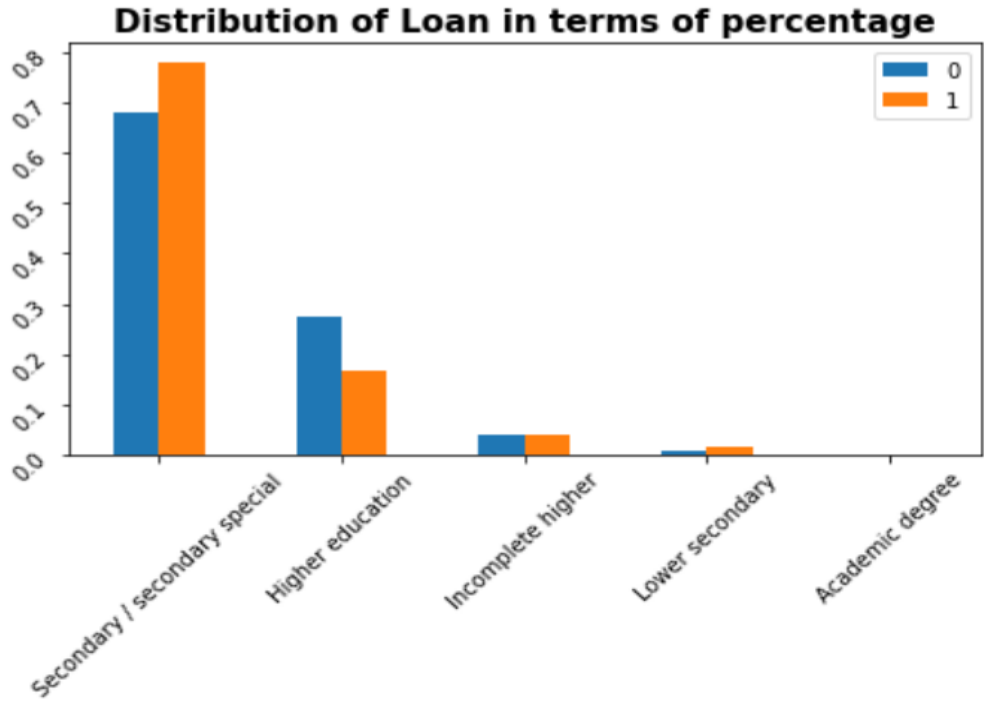
- As per pi-chart we can conclude that the imbalance between Defaulter & Non-Defaulter TARGET variables is high .
- The Non-Defaulters Percentage are 91%
- Whereas the Defaulters are 9%

UNIVARIATE ANALYSIS ON NEW APPLICATION DATA SET

➢ Univariate Categorical Ordered Analysis



CODE_GENDER



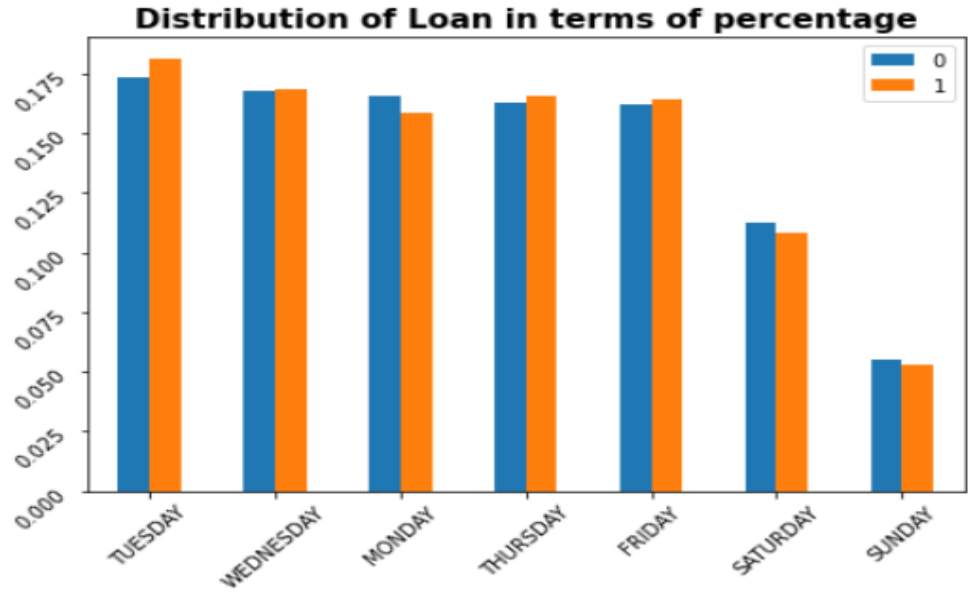
NAME_EDUCATION_TYPE

❖ **Gender-**

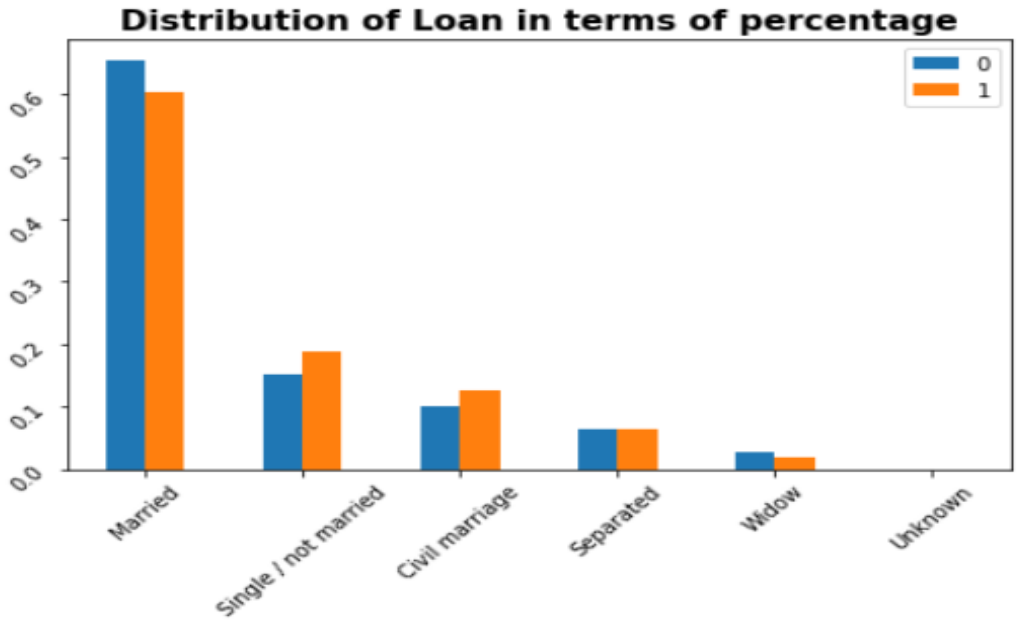
- Less number of males take loans but the defaulters are higher in case of males.
- Comaparing both the plots for Payment Difficulties and Non Payment Difficulties on the basis of Code_Gender, we observe that there is an increase in the percentage in Male Payment Difficulties for Non-Defaulters plot although Females percentage are the majority in both the cases.

❖ **Education-**

- The default rate in secondary education is much high and for higher education is much low.
- Most customers take loan for secondary education followed by higher education.



NAME_HOUSING_TYPE

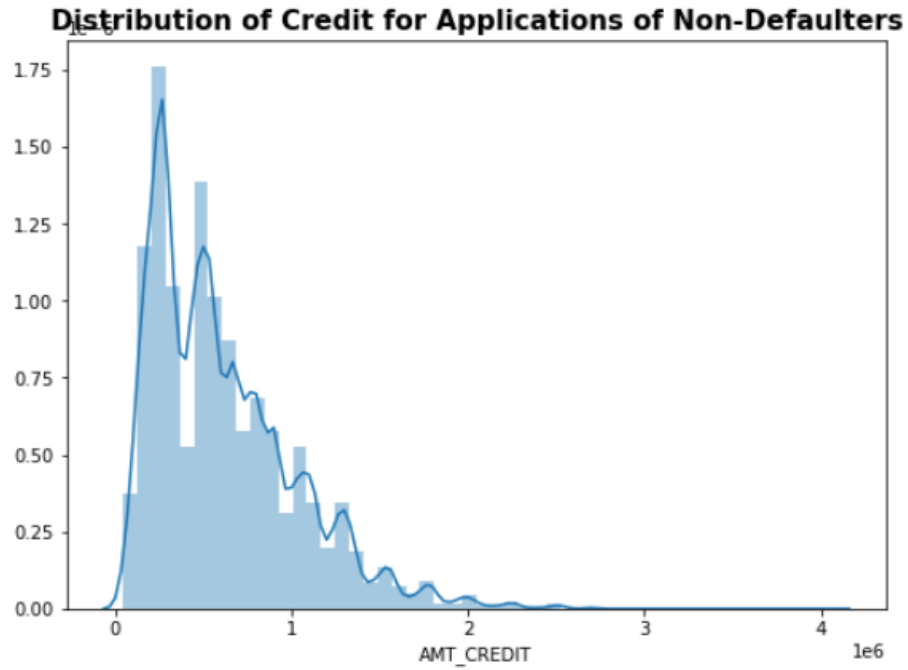


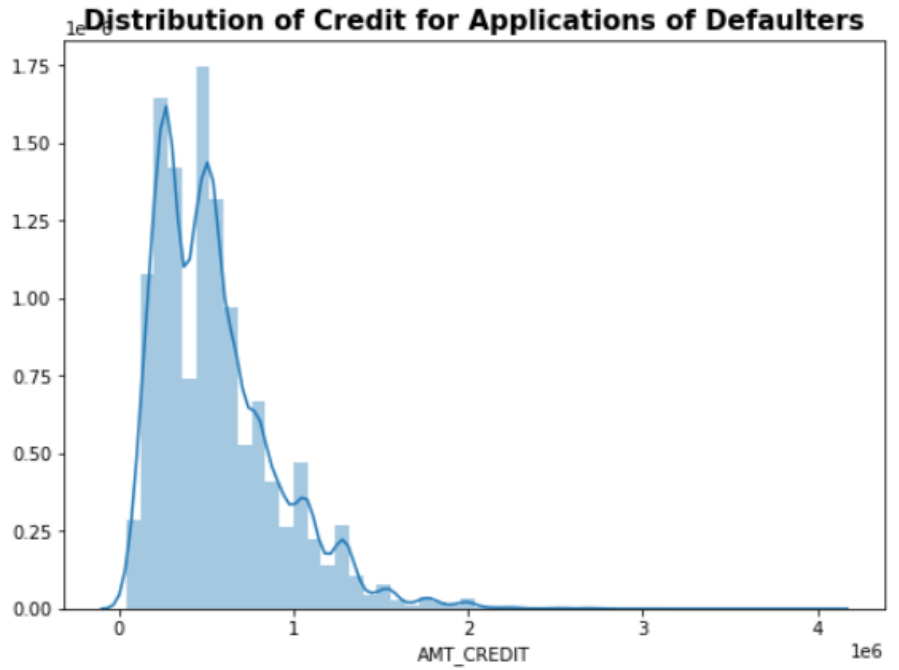
NAME_FAMILY_STATUS

Distribution of Housing Type and Family Status:

- Housing Type-
  - We observe People living with parents tend to default more often when compared with others.
  - We observe an increase in the percentage of Payment Difficulties who live with their parents when compared to the percentages of Defaulters and Non-Defaulters cases.
  - It is clear from the graph that people who have House/Appartment, tend to apply for more loans.
- Family Status-
  - We can see observe that the Increase in percentage of Civil and Single/not married with Loan-payment difficulties (Defaulters case) as compare to Non-Defaulters Case.
  - We can see observe that the Decrease in percentage of Married and Widow with Loan-payment difficulties (Defaulters case) as compare to Non-Defaulters Case.

➢ UNIVARIATE CONTINUOUS VARIABLE ANALYSIS





❖ **Proportion of Defaulters by Credit amount-**

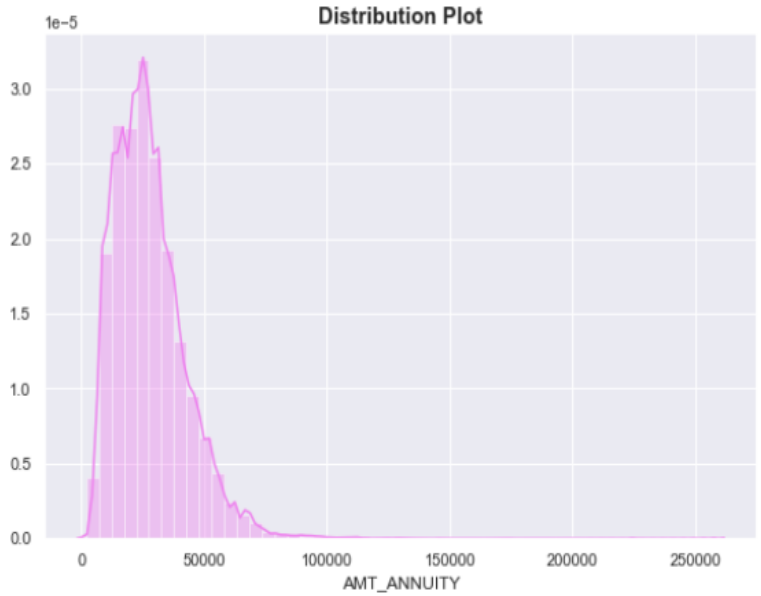
- From above plot we can conclude that better return of Loans as loan size increases.
- There doesn't seem to be a clear distiguish between the group which defaulted vs the group which didn’t.

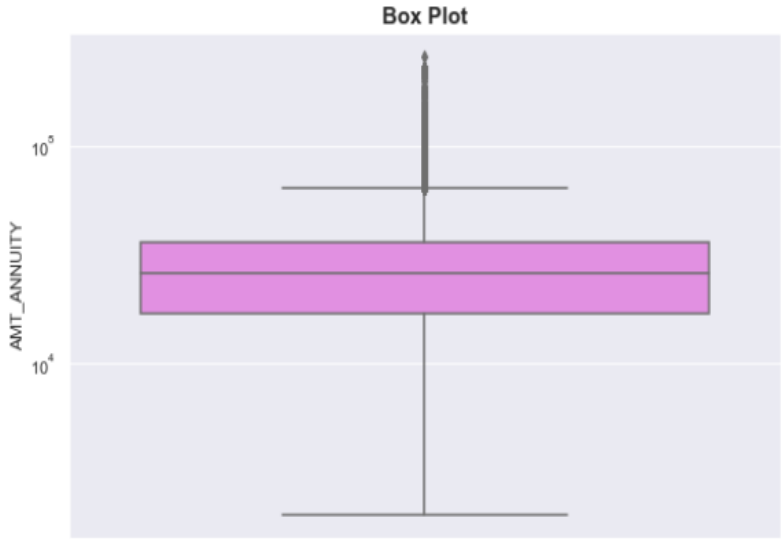
➢ UNIVARIATE ANALYSIS OF NUMERICAL VARIABLES ON THE BASIS OF 'TARGET' VARIABLE

❖Loan Annuity-

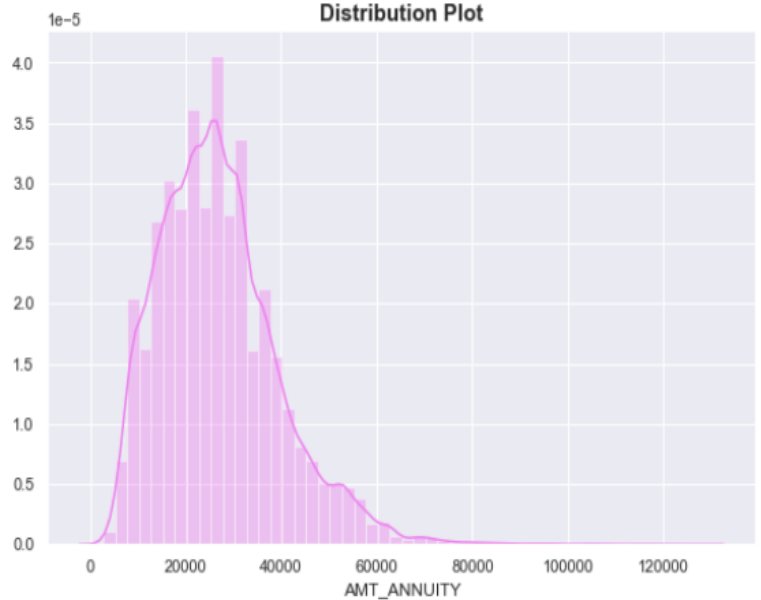
For Target = 0

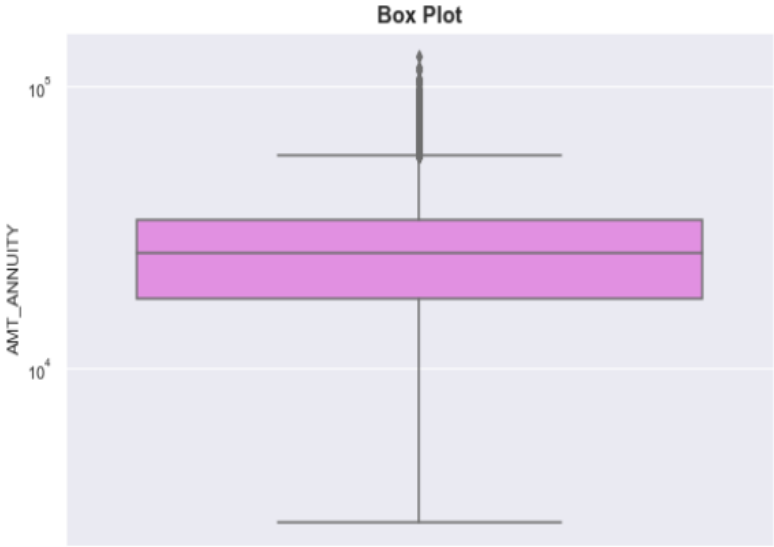
- We can observe some outliers and the first quartile is bigger than third quartile for annuity amount which means most of the annuity clients are from first quartile.





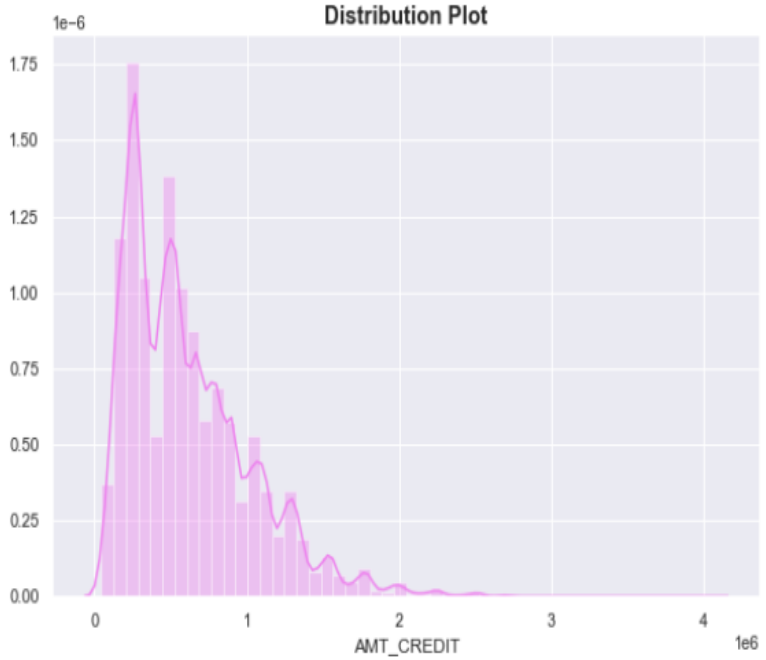
For Target = 1

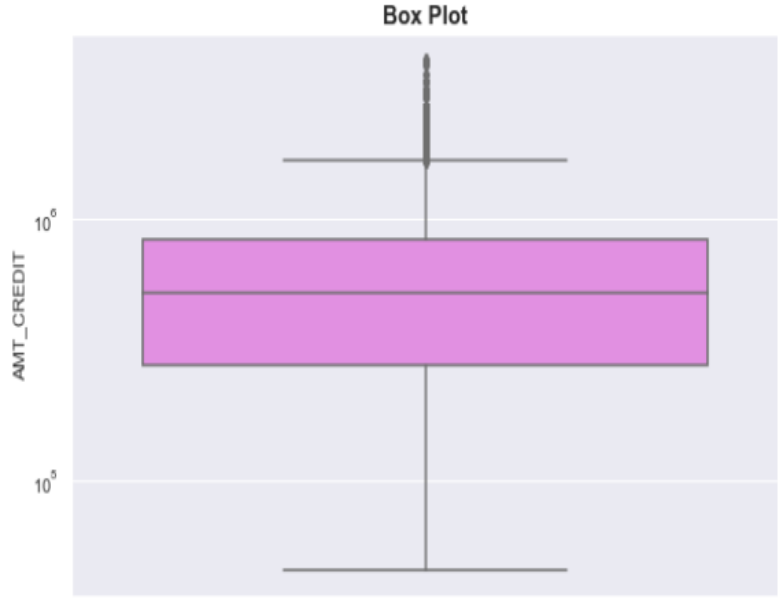




➢ Credit Amount-

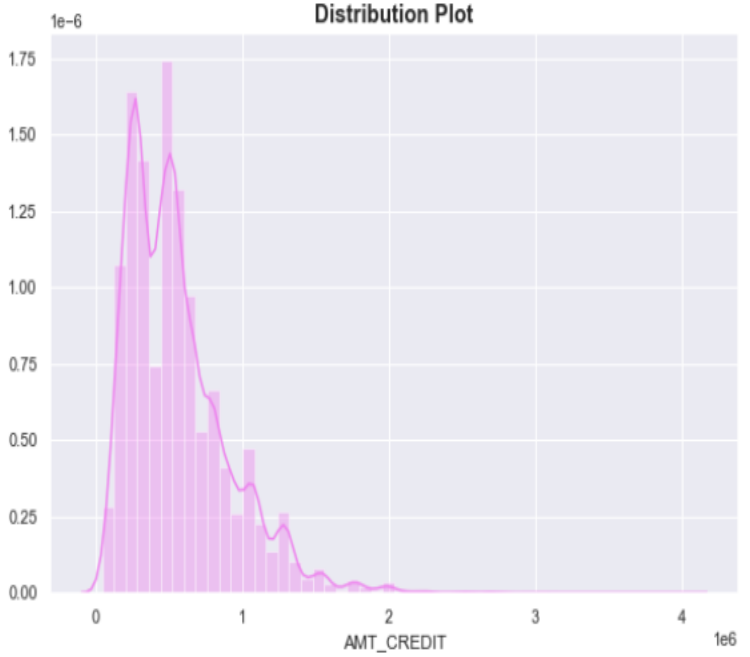
Target = 0

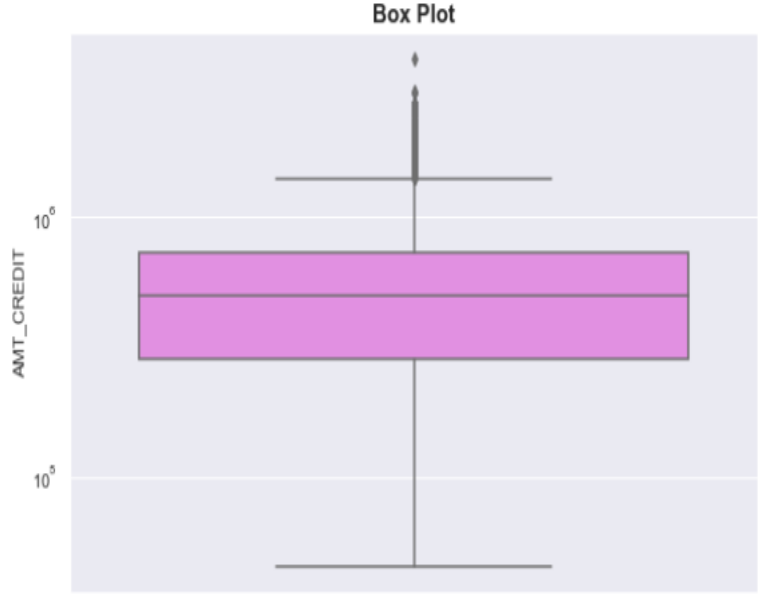




- We can observe that some outliers and the first quartile is bigger than third quartile for annuity amount.
- Most of the annuity clients are from first quartile.

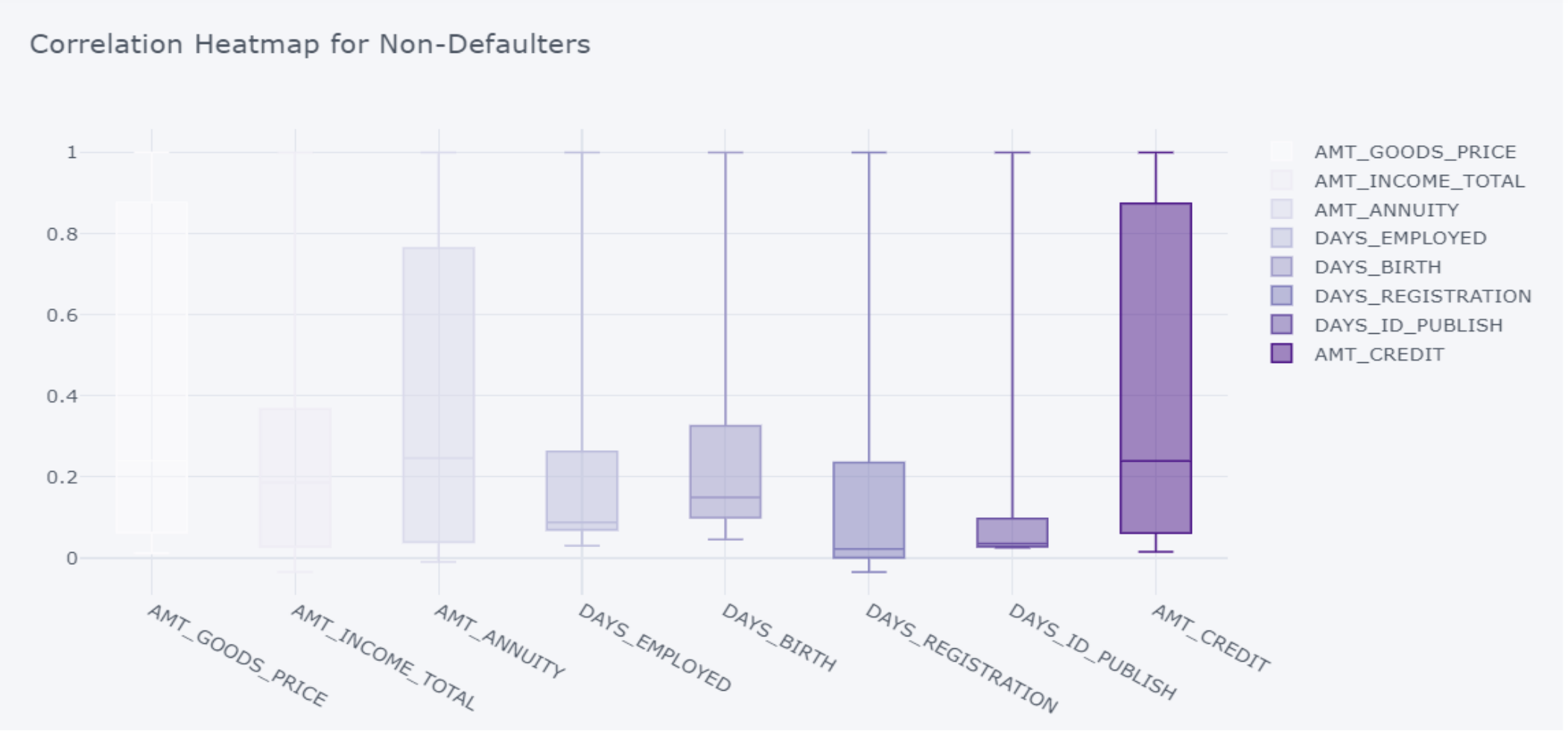
Target = 1



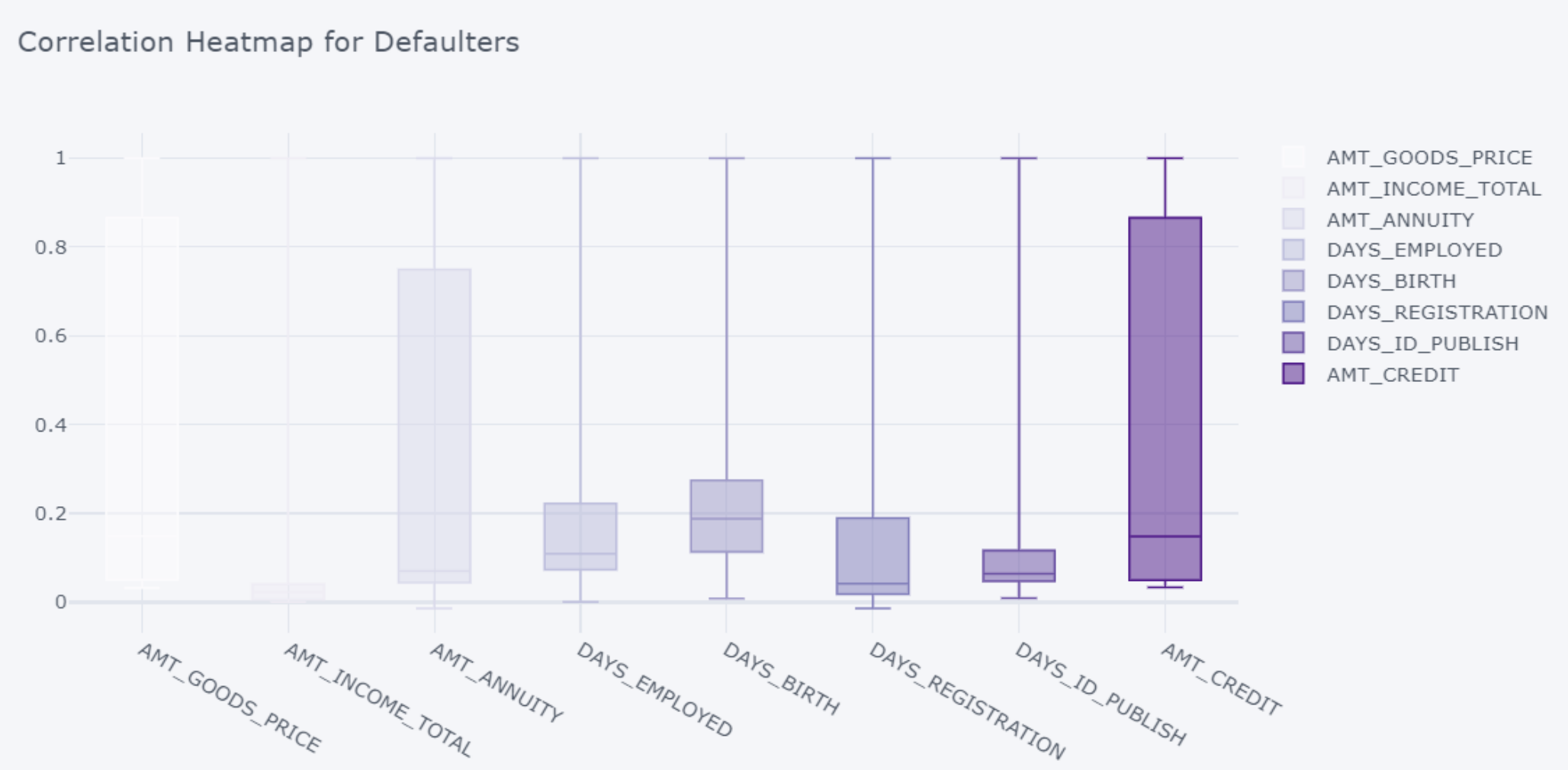


➢ BIVARIATE ANALYSIS OF NUMERICAL VS NUMERICAL VARIABLES

CORRELATION HEATMAP FOR NON-DEFAULTERS



CORRELATION HEATMAP FOR NON-DEFAULTERS



TOP 10 CORRELATIONS FOR NON DEFAULTERS AND DEFAULTERS

	Column1	Column2	Correlation	Abs_Correlation
1268	AGE_YEARS	DAYS_BIRTH	0.999591	0.999591
960	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998492	0.998492
220	AMT_GOODS_PRICE	AMT_CREDIT	0.986471	0.986471
517	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.949504	0.949504
434	CNT_FAM_MEMBERS	CNT_CHILDREN	0.893275	0.893275
997	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.861492	0.861492
665	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.860421	0.860421
776	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.820828	0.820828
221	AMT_GOODS_PRICE	AMT_ANNUITY	0.766655	0.766655
184	AMT_ANNUITY	AMT_CREDIT	0.762103	0.762103

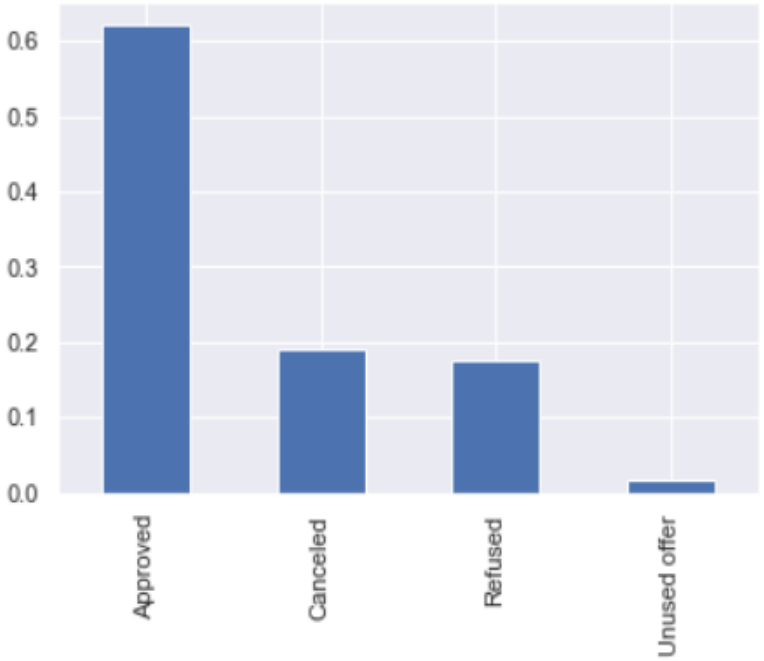
	Column1	Column2	Correlation	Abs_Correlation
1268	AGE_YEARS	DAYS_BIRTH	0.999565	0.999565
960	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998289	0.998289
220	AMT_GOODS_PRICE	AMT_CREDIT	0.982464	0.982464
517	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956531	0.956531
434	CNT_FAM_MEMBERS	CNT_CHILDREN	0.893829	0.893829
997	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.867983	0.867983
665	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.846872	0.846872
776	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.768247	0.768247
221	AMT_GOODS_PRICE	AMT_ANNUITY	0.748940	0.748940
184	AMT_ANNUITY	AMT_CREDIT	0.748708	0.748708

- We observe that there is a high correlation between credit amount and goods price.
- High similarities between both the correlations.

➢ UNIVARIATE ANALYSIS ON NEW APPLICATION DATA SET

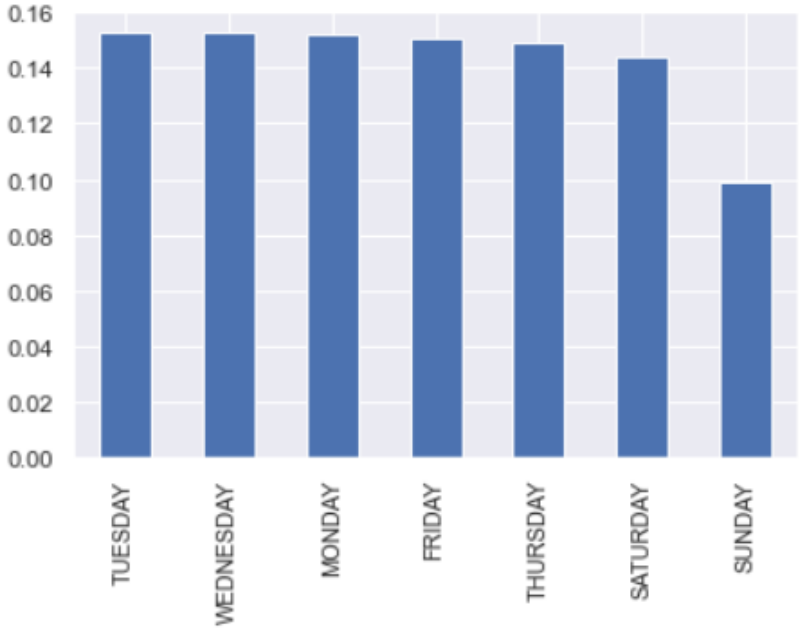
UNIVARIATE ANALYSIS ON CATEGORICAL COLUMN

* Contract Status



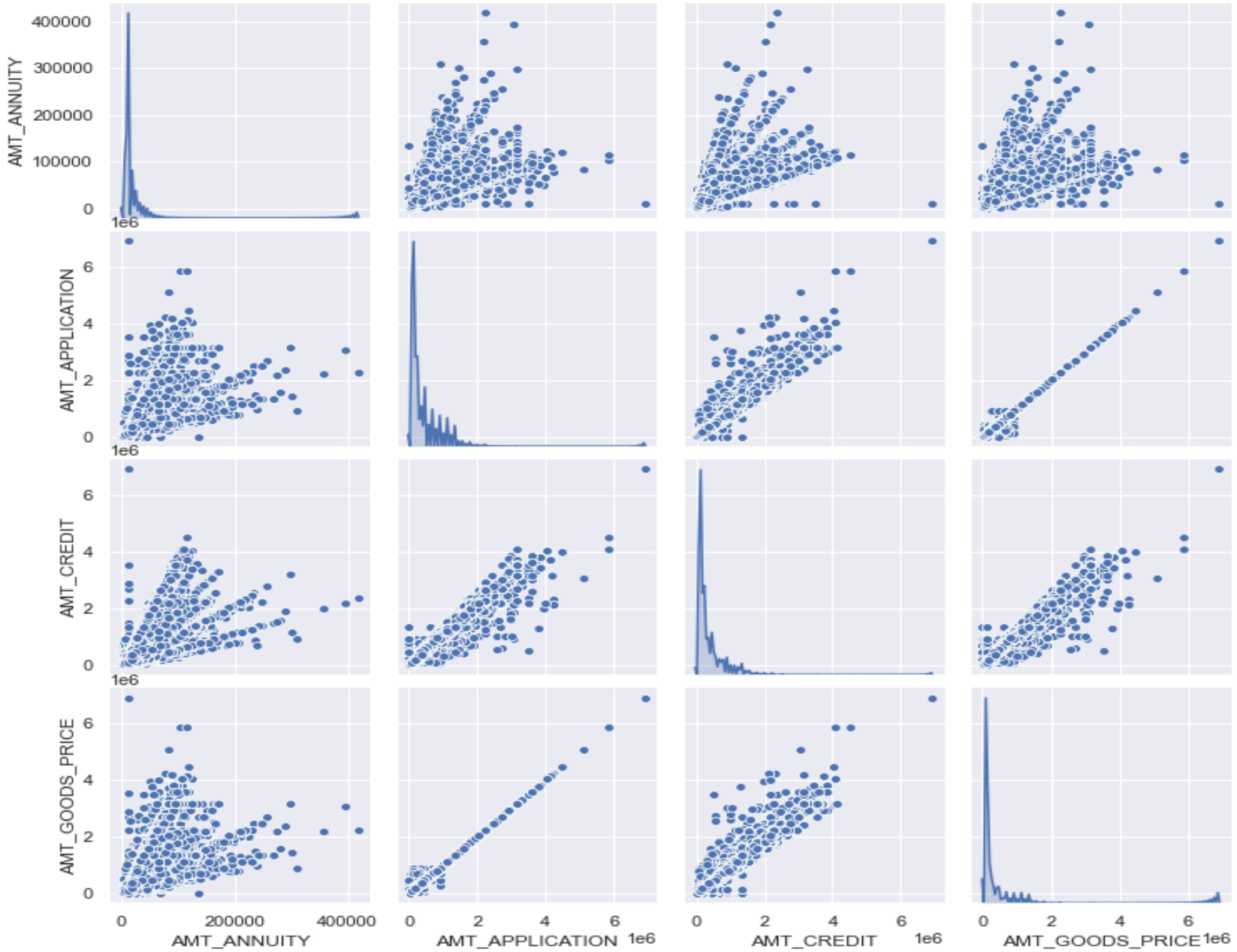
- Here the majority of loans are approved and very less percentage of loans are unused offer

* Day of the week



- Less number of applicants that come in the weekends.

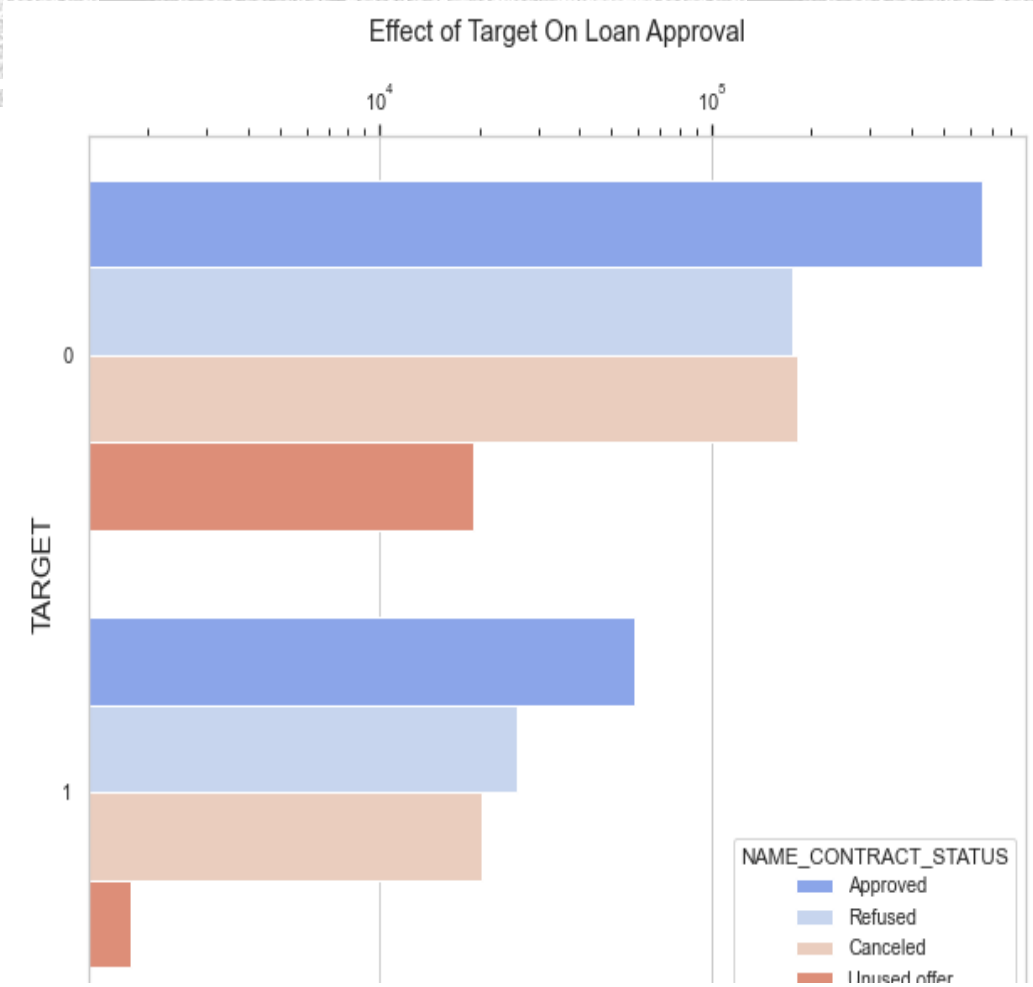
➢ BIVARIATE ANALYSIS ON NUMERICAL COLUMNS



❖ FEW POINTS CAN BE CONCLUDED FROM THE PAIRPLOT :

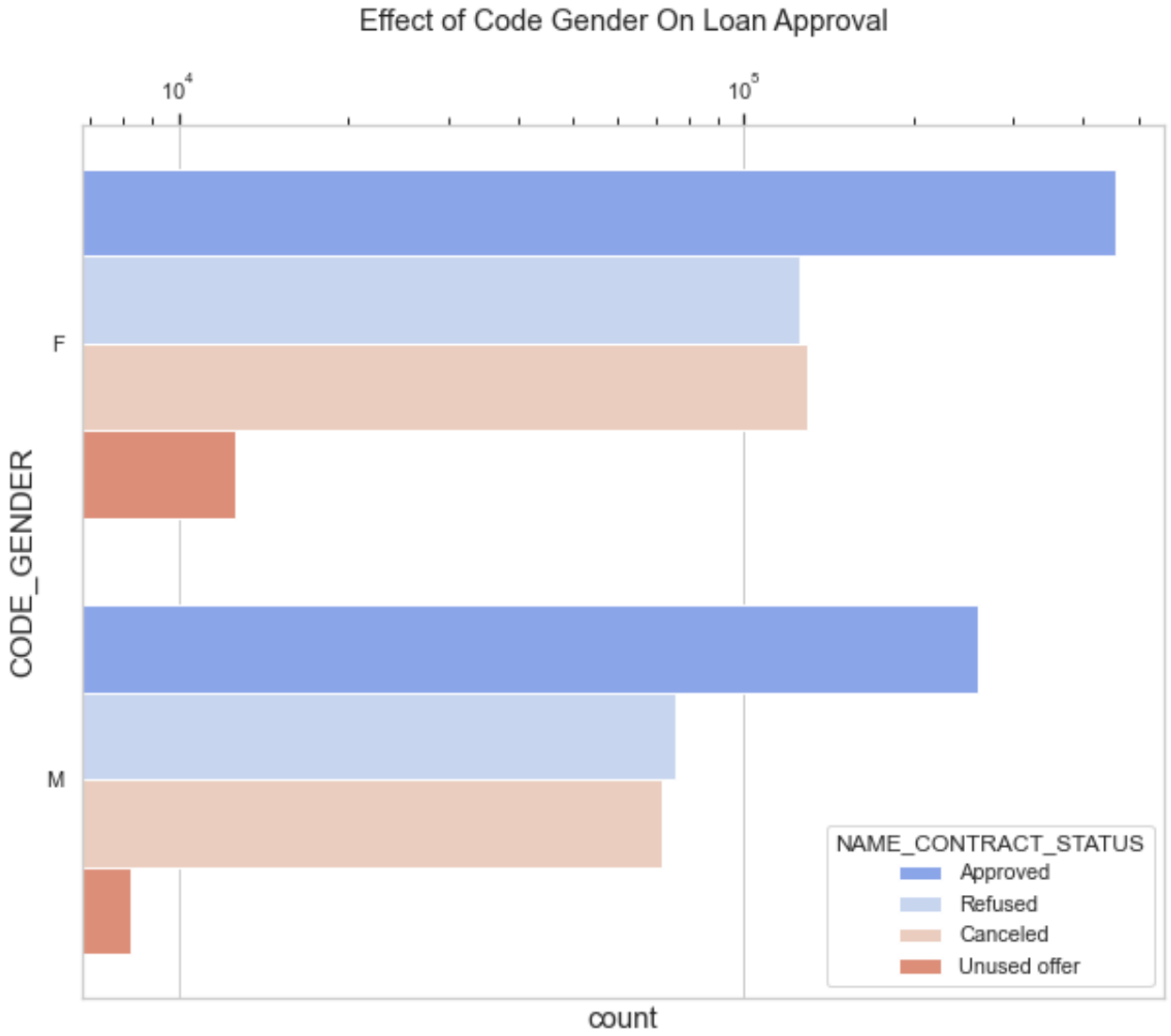
- Column CNT_Payment ideally should have had a high correlation with AMT_credit, i.e. higher credit, more the term of loan. But no such correaltion can be seen.
- AMT_GOODS_PRICE, AMT_ANNUITY, AMT_APPLICATION - as expected have high correlation.
- Higher the value of good purchased more there will be need of loan and surely all these will correlate.
- AMT_Credit to AMT_GOOD_PRICE also the correlation is high.

➢ PERFORMING UNIVARIATE ANALYSIS ON MERGED DATA SET



- **Distribution of Contract Status with Target-**
- Loans which were previously refused or cancelled have a higher default rate.

➢ PERFORMING UNIVARIATE ANALYSIS ON MERGED DATA SET

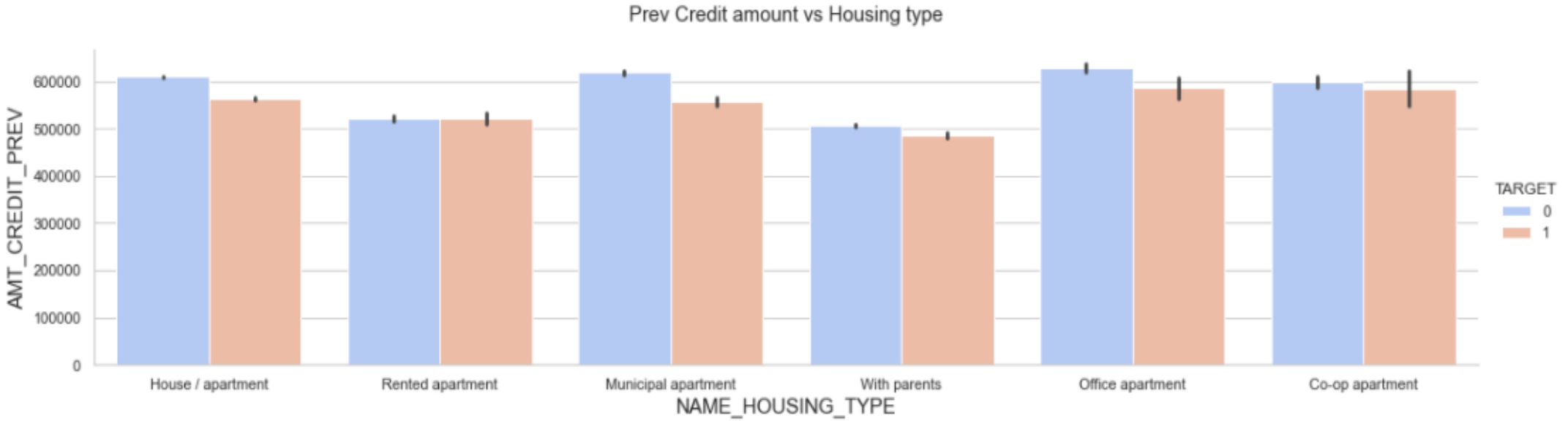


❖ **Effect of Code Gender on Loan Approval-**

- As per above plot, We can see that code gender doesn't have any effect on application approval or rejection.

➢ PERFORMING BIVARIATE ANALYSIS ON MERGED DATA SET

AMT_CREDIT_PREV (CREDIT AMOUNT PREV) VS NAME_HOUSING_TYPE (HOUSING TYPE)



- we can conclude that bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment.

CONCLUSION

EDA for Banking Data Sets revealed that:

- - The Bank lends more to females.
- - The Proportion of defaulters is 9%.
- - Proportion of working defaults more and state servants less.
- - Banks should be less focused on income type 'Working' as they are having large number of unsuccessful payments.
- - Higher amount loans , Higher Income Less Defaults.
- - Banks should be more focused on contract type 'Student' ,'pensioner' and 'Businessman' with housing 'type other than 'Co-op apartment' for successful payments.
- Get as much as clients from housing type 'With parents' as they are having least number of unsuccessful payments.
- - 'Repair' is having higher number of unsuccessful payments on time with loan purpose.
- - Loans previously refused/cancelled - Higher the Default Rate.