EXPLORATORY DATA ANALYSIS – AVINASH KUMAR

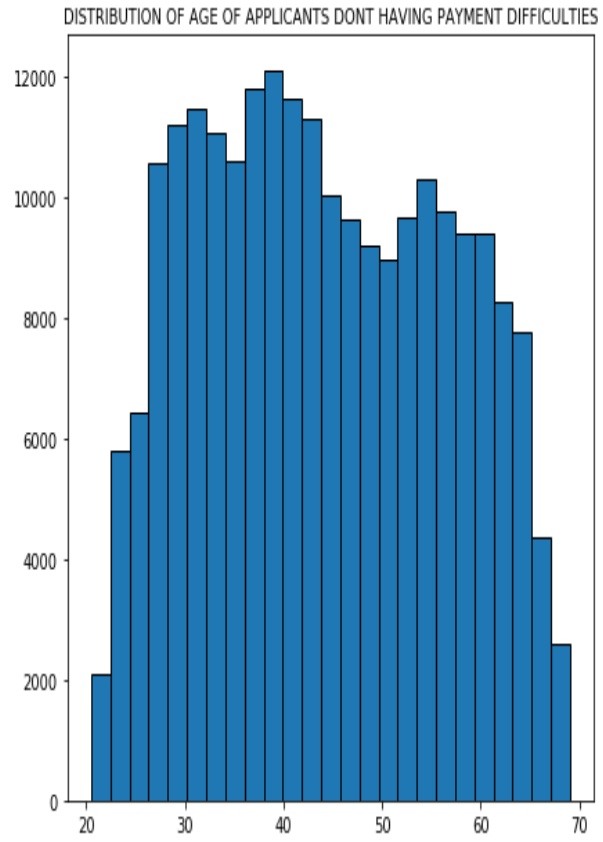
NITISH RANJAN JHA

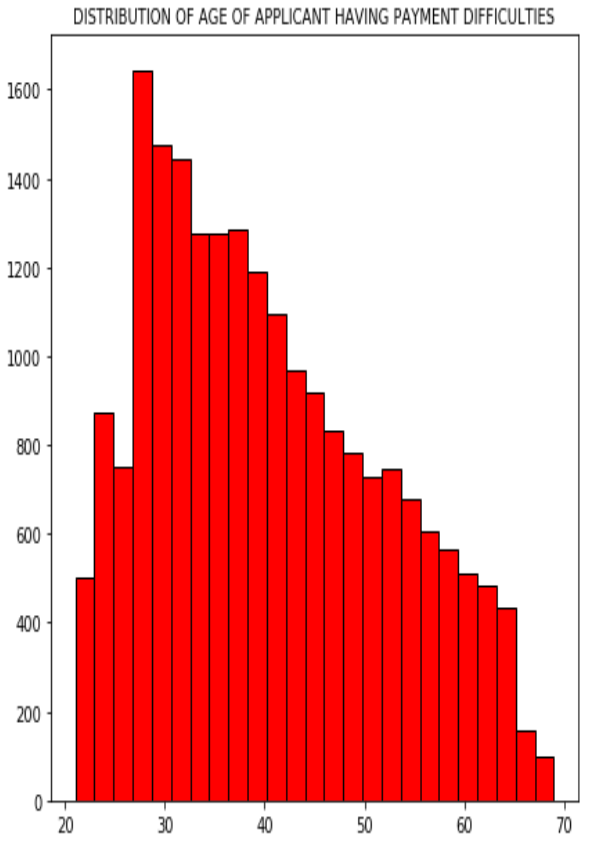
EFFECT OF AGE ON TARGET

1.We found that percentage of being a defaulter tend to decrease as getting older.

2.Data provided to us covers the age group from 20 to 70 years.

3.Clients who lie in Age Group 20-25 were 12% more likely to be defaulter and least were in age group 65- 70 with approx. 3%.



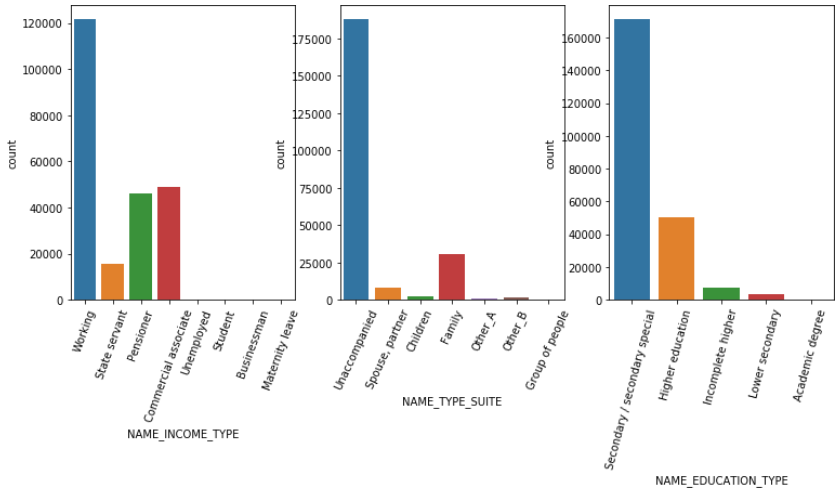


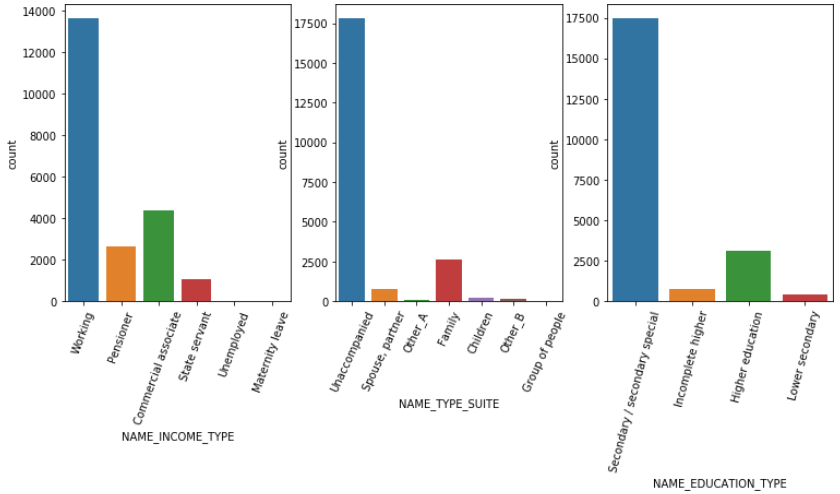
CLIENTS LIKELY TO BE DEFAULTERS :

1.Upper plot is for those folks(who don't have any difficulties paying for loans) and below plot is for the folks who likely to be Defaulters

2.We can conclude that income type group like student and businessman are less likely to be defaulters

3.Meanwhile clients with academic scholar as Education degree are less likely to be Defaulter.



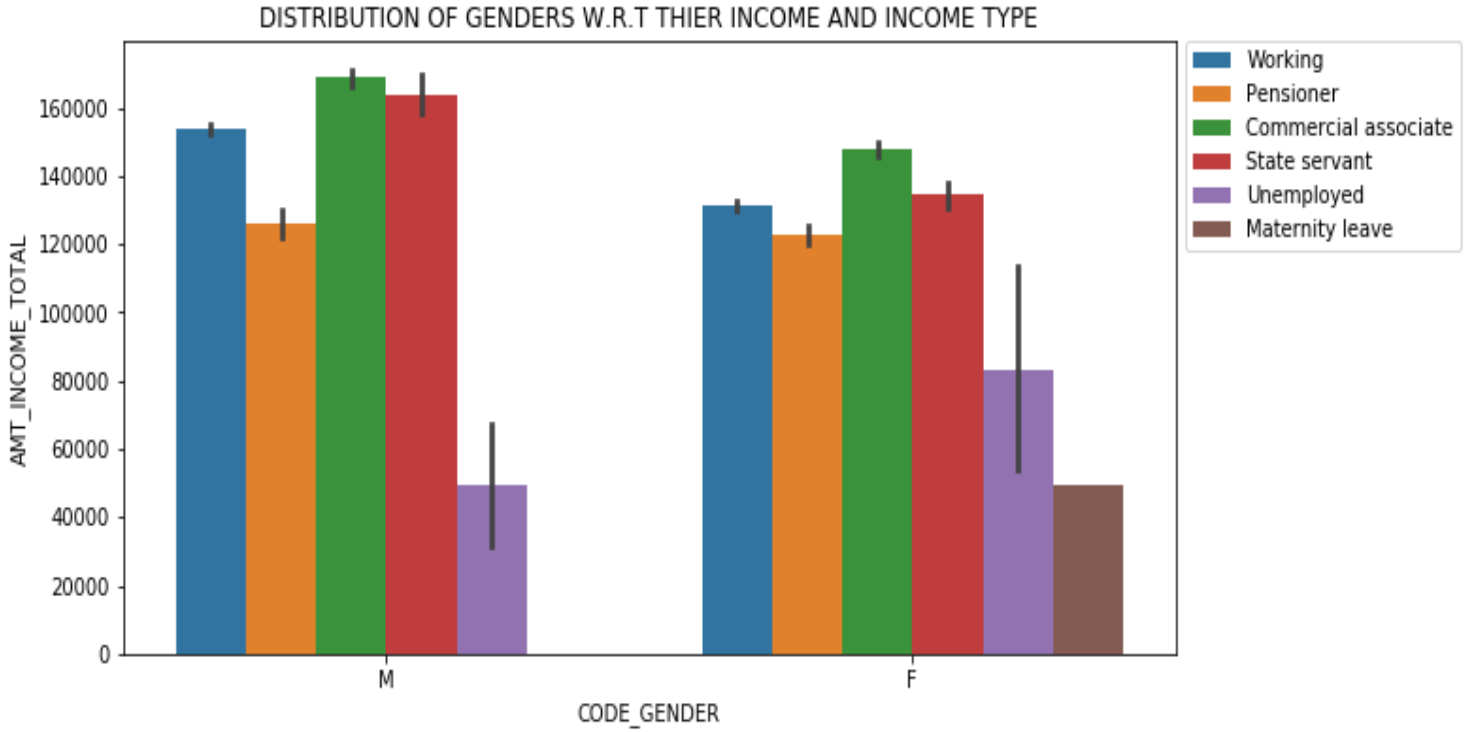


EFFECT OF GENDER , INCOME ON TARGET

I.Plot represent the distribution of genders w.r.t their income and income type:

II.Top 3 likely to be defaulters on the basis of income type with respect to gender and total income is

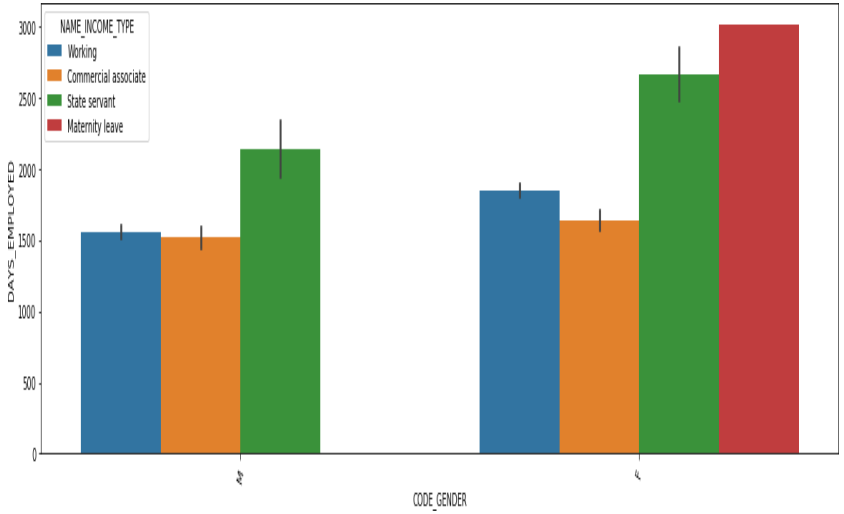
1. Commercial Associate
2. State Servant
3. Working client

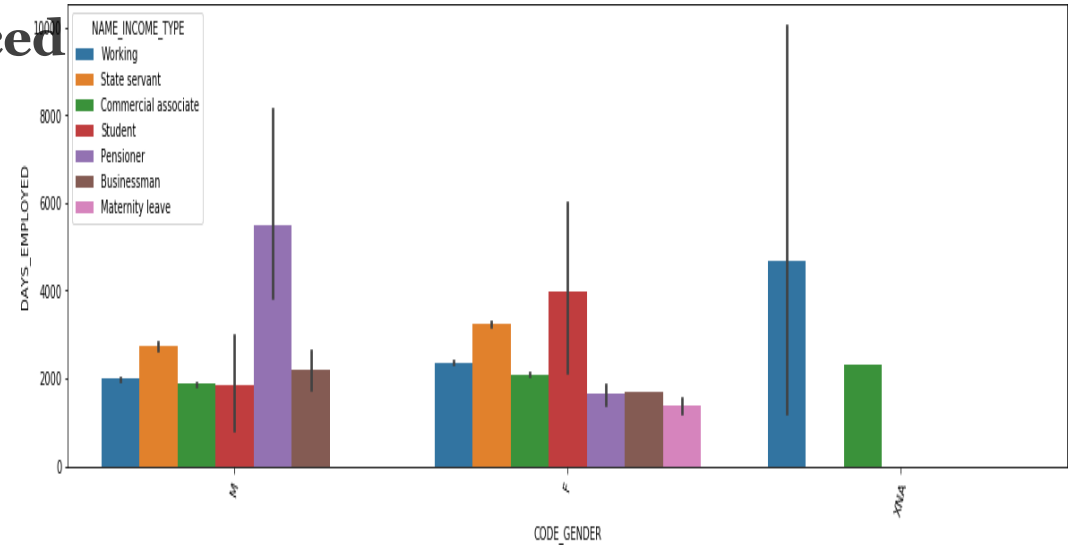


Effect of Gender , Days employed and income type together on Target

1.from plots we can conclude that state servent(male and female) and Female on maternity leave are most experienced in case of defaulters.

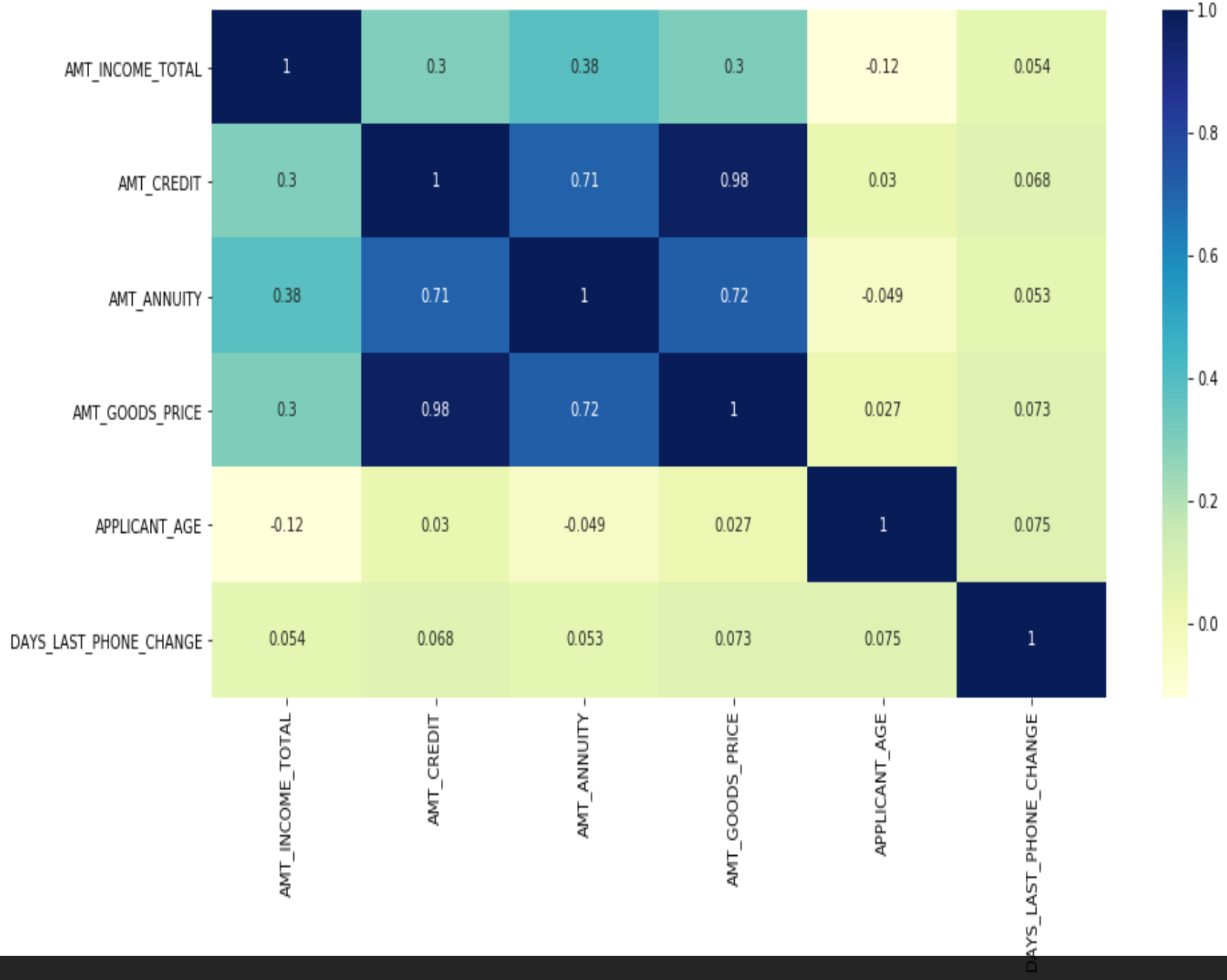
2.Meanwhile for non defaulters , most experienced are pensioner in males and student in female.





CORRELATION MATRIX OF THE VARIABLES WHO ARE LIKELY TO **PAY** ON TIME

From the correlation matrix we can inference that the correlation between amount of goods price and credit is very high which inference that the folks who have applied for consumer loans are likely to default more than the folks that have applied for cash loan.
We can also see that there is almost no effect of applicant age and last_day _phone_change on being defaulter.

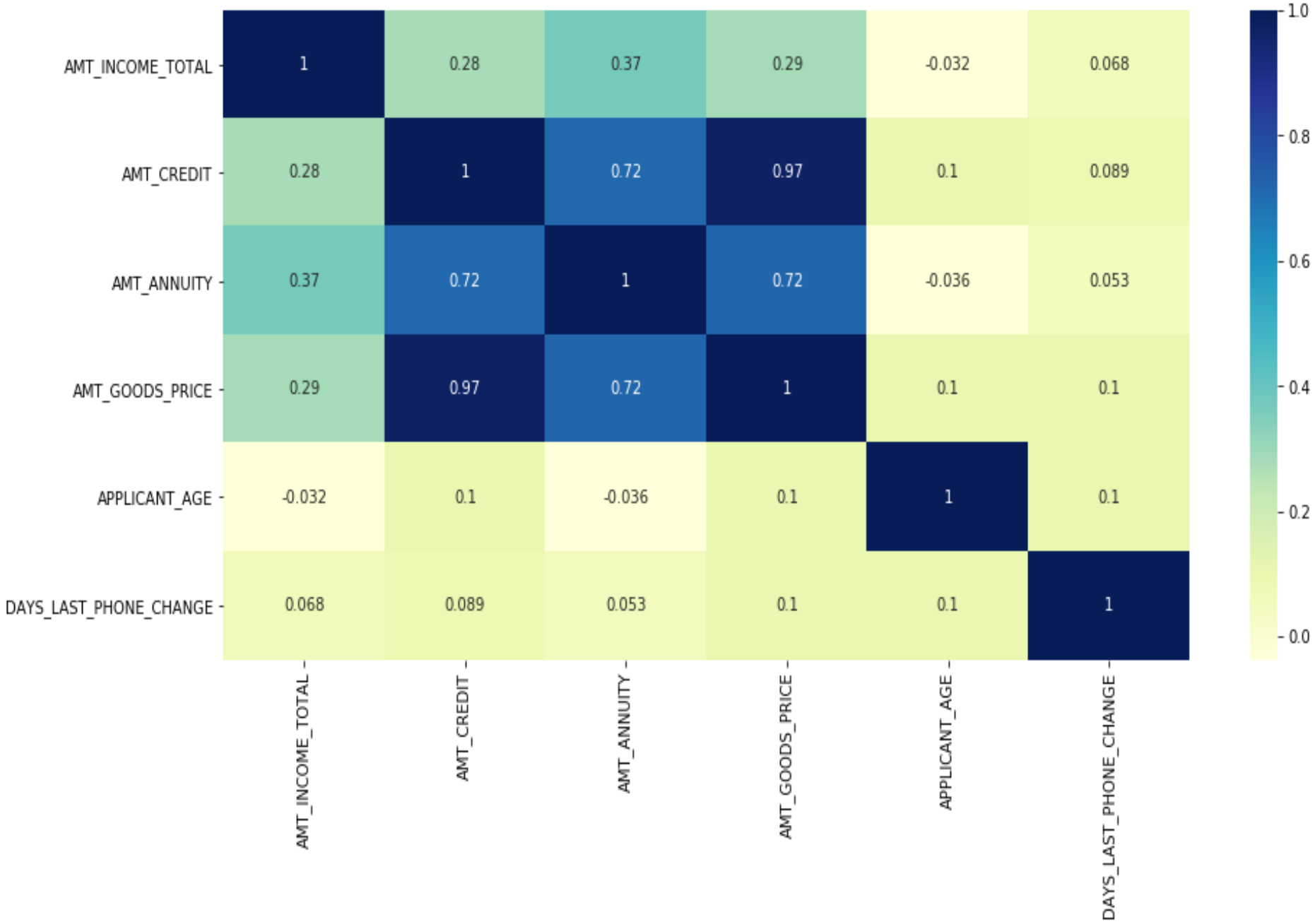


TOP-10 CORRELATION WHICH HAVE AN HIGHER POSSIBILITIES TO **NOT** BE A DEFAULTER AND MUCH LIKELY TO PAY ON TIME FROM NEW APPLICATION DATASET

	Variable_1	Variable_2	CORR
19	AMT_GOODS_PRICE	AMT_CREDIT	0.977363
20	AMT_GOODS_PRICE	AMT_ANNUITY	0.720037
13	AMT_ANNUITY	AMT_CREDIT	0.711065
12	AMT_ANNUITY	AMT_INCOME_TOTAL	0.384770
18	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.302398
6	AMT_CREDIT	AMT_INCOME_TOTAL	0.295562
24	APPLICANT_AGE	AMT_INCOME_TOTAL	0.120130
34	DAYS_LAST_PHONE_CHANGE	APPLICANT_AGE	0.074674
33	DAYS_LAST_PHONE_CHANGE	AMT_GOODS_PRICE	0.072964
31	DAYS_LAST_PHONE_CHANGE	AMT_CREDIT	0.068039

CORRELATION MATRIX OF THE VARIABLES WHO ARE LIKELY TO BE A **DEFAULTER**

From the correlation matrix we can inference that the correlation between amount of goods price and credit is very high which inference that the folks who have applied for consumer loans are likely to default more than the folks that have applied for cash loan.
We can also see that there is almost no effect of applicant age and last_day _phone_change on being defaulter.



TOP-10 CORRELATION WHICH HAVE AN HIGHER POSSIBILITIES TO BE A **DEFAULTER** AND NOT MUCH LIKELY TO PAY ON TIME FROM NEW APPLICATION DATASET

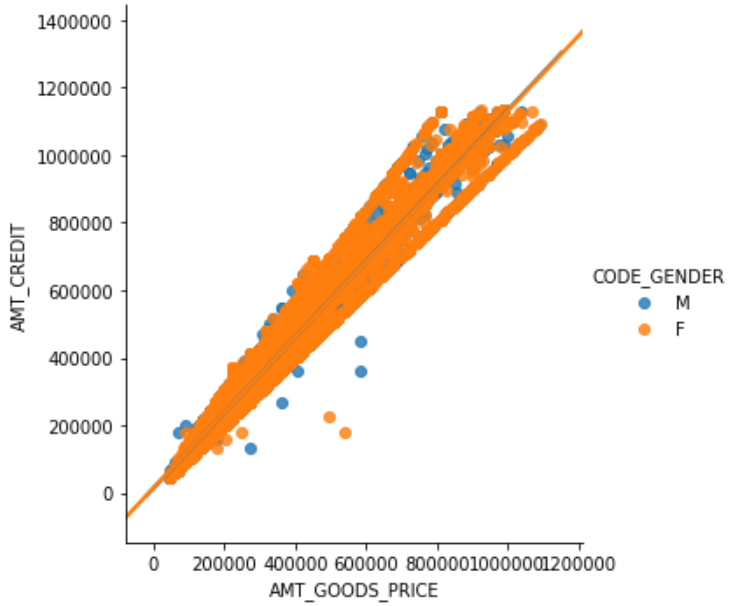
	Variable_1	Variable_2	CORR
19	AMT_GOODS_PRICE	AMT_CREDIT	0.972356
20	AMT_GOODS_PRICE	AMT_ANNUITY	0.723348
13	AMT_ANNUITY	AMT_CREDIT	0.716199
12	AMT_ANNUITY	AMT_INCOME_TOTAL	0.369497
18	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.286447
6	AMT_CREDIT	AMT_INCOME_TOTAL	0.279945
25	APPLICANT_AGE	AMT_CREDIT	0.104597
34	DAYS_LAST_PHONE_CHANGE	APPLICANT_AGE	0.104036
27	APPLICANT_AGE	AMT_GOODS_PRICE	0.103487
33	DAYS_LAST_PHONE_CHANGE	AMT_GOODS_PRICE	0.102389

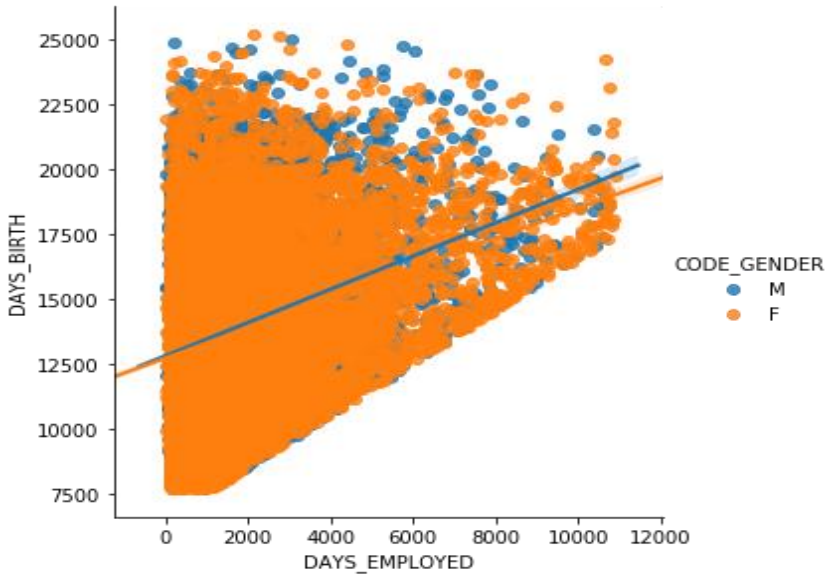
TOP CORRELATION VARIABLE

1)Column-1 represent lm plots represent the correlation between amount credit and amount of good price of the people who don't have difficulty in loan repayment

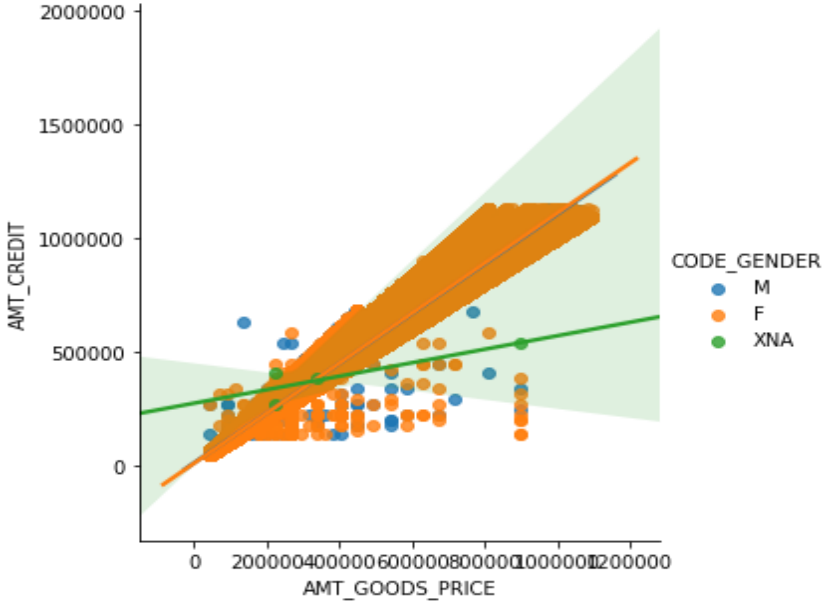
2)Column-2 represent lm plots represent the correlation between amount credit and amount of good price of the people who have difficulty in loan repayment

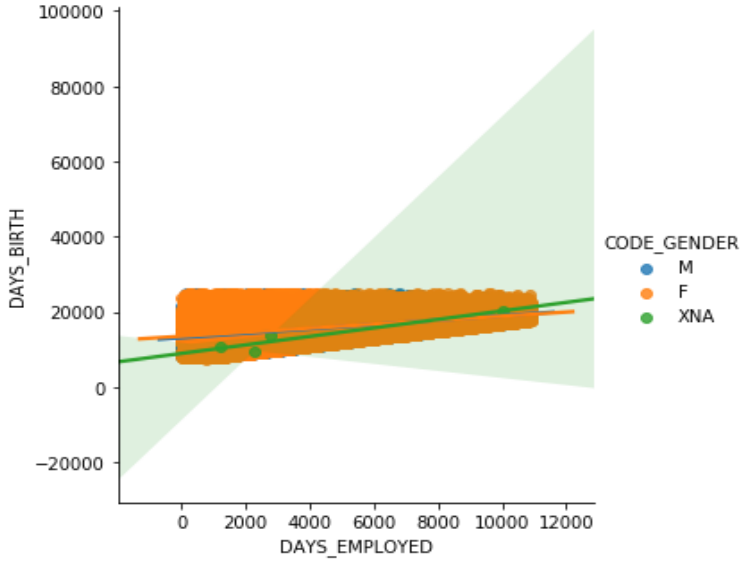
COLUMN-1





COLUMN-2



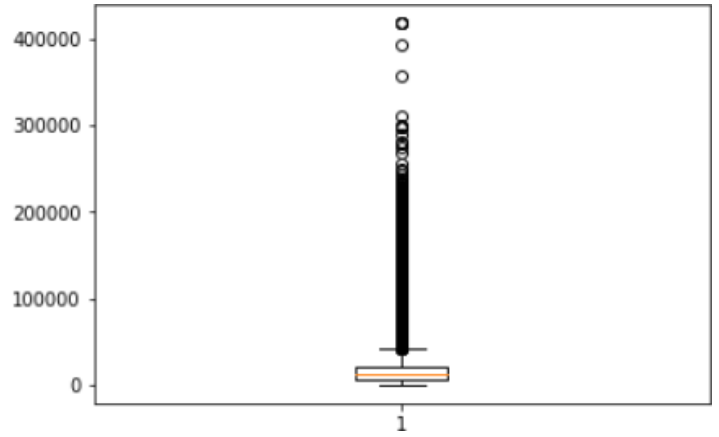


PreviousApplication

ANALYSIS

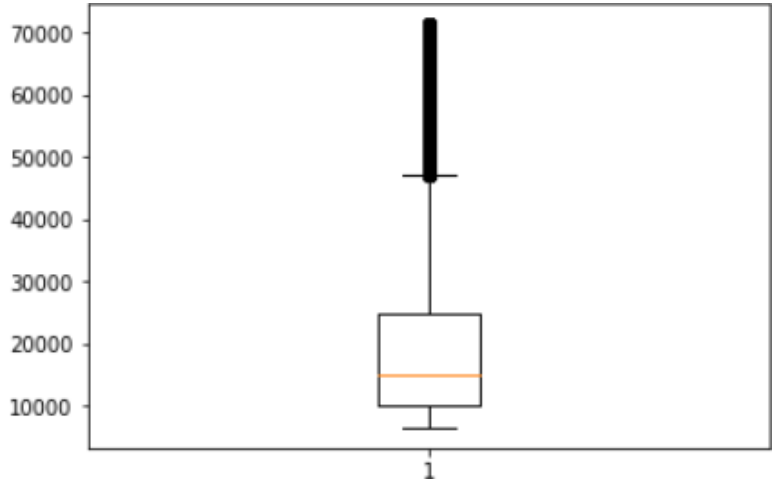
OUTLIER'S TREARMENT

For this , we used ANNUITY AMOUNT column as the base column for further analysis , and after doing the same we found that there are outliers present in significant amount , which can be seen in plot no-1



To remove the outliers, we have divided the data into different percentiles:5%,10%,25%,50%,75%,85%,90%,95%,99%

The first quartile of data was holding low values in significant amount, hence to remove the outliers we have selected a range between 25th and 95th percentile:

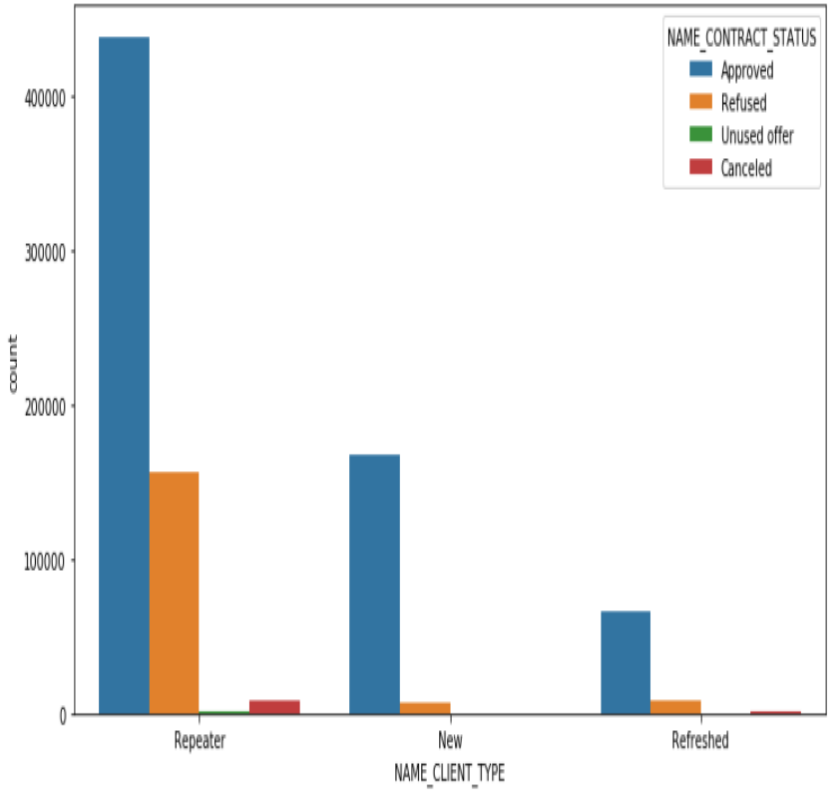


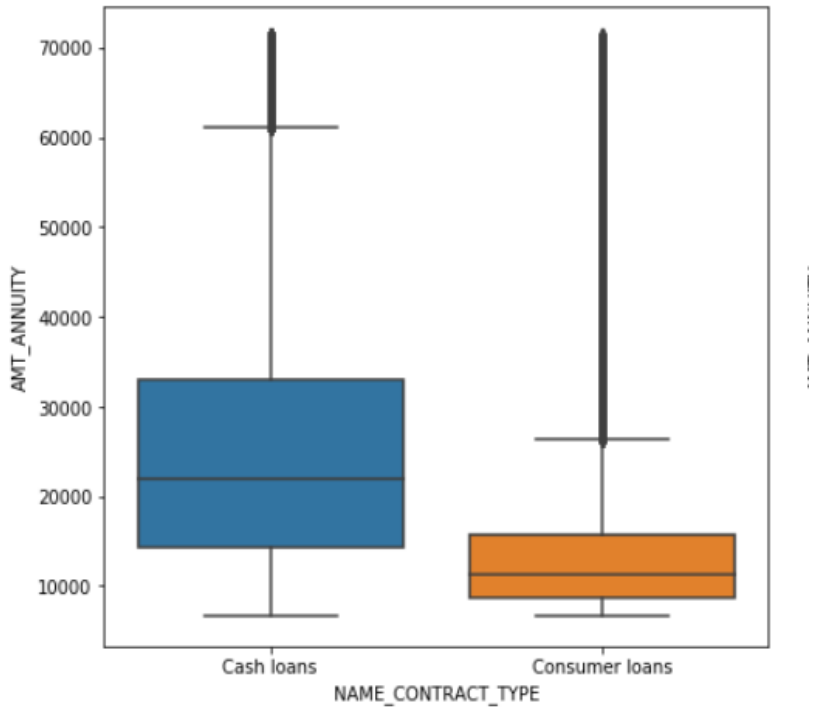
EXPLORATORY DATA ANALYSIS

From the analysis we have found the following points-

1) The cash loans are the main driver, however the consumer loans have higher annuity amount:

2)Rejection rate is high for the repeated customers:

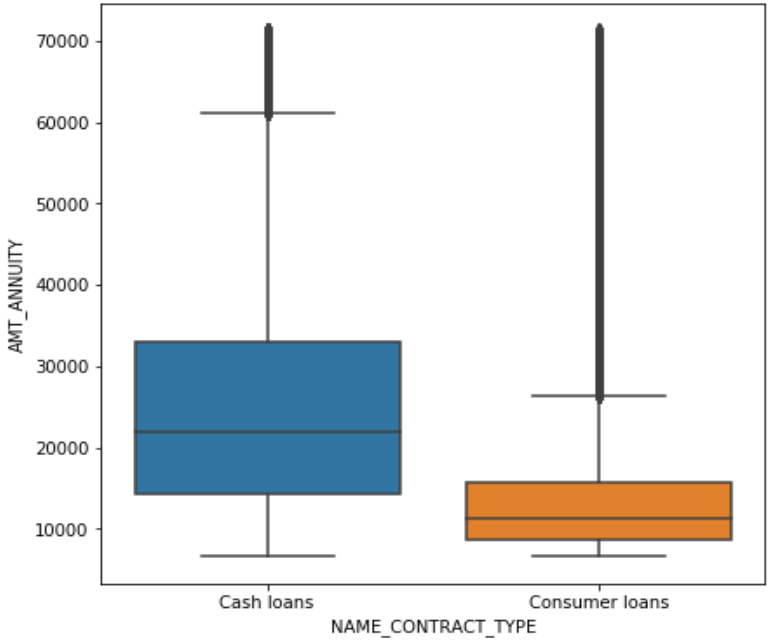


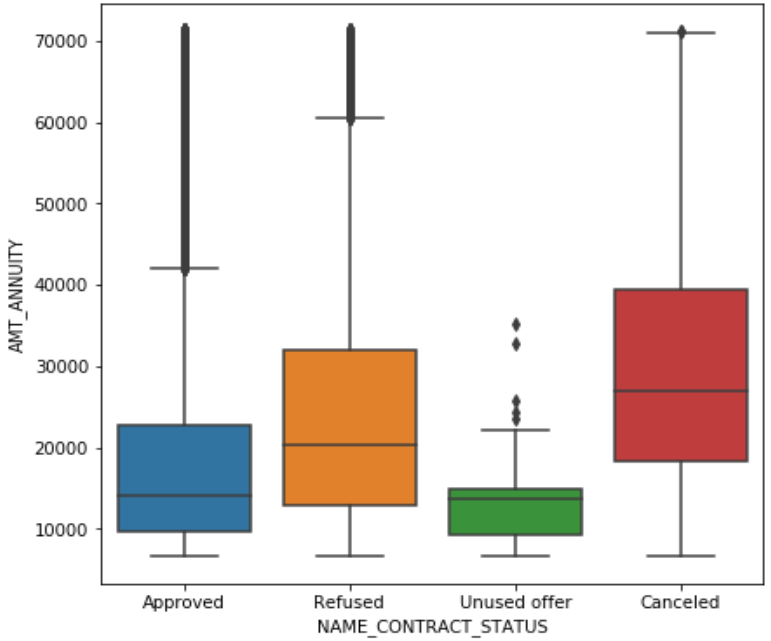


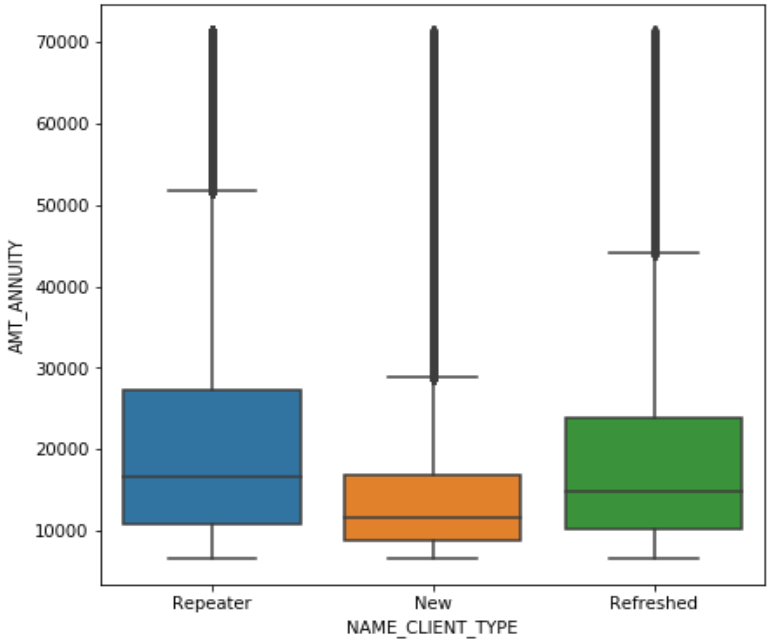
From the above graph we can inference following-

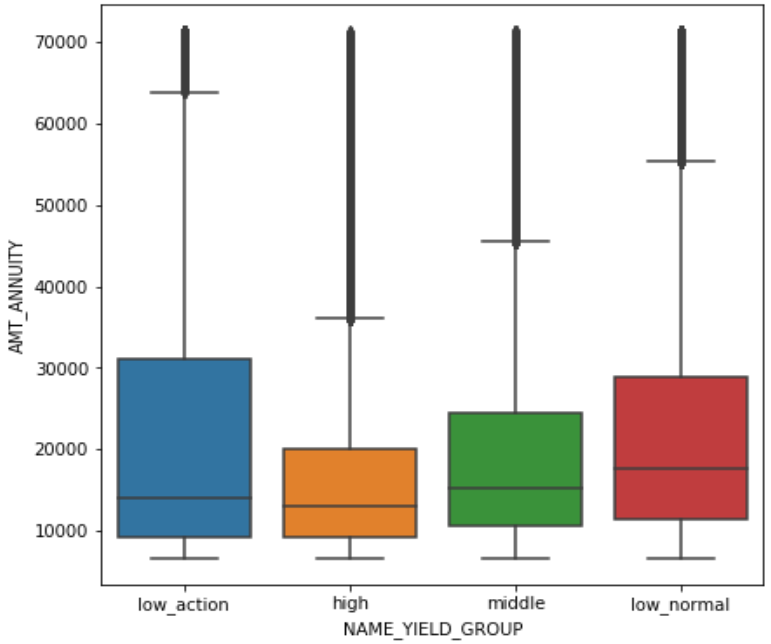
1)Cash loans are the main driver and however the consumer loans have higher annuity amount.

2)Clients tends to repeat their application but for lower annuity amount as the fresh clients tends to have higher annuity amount



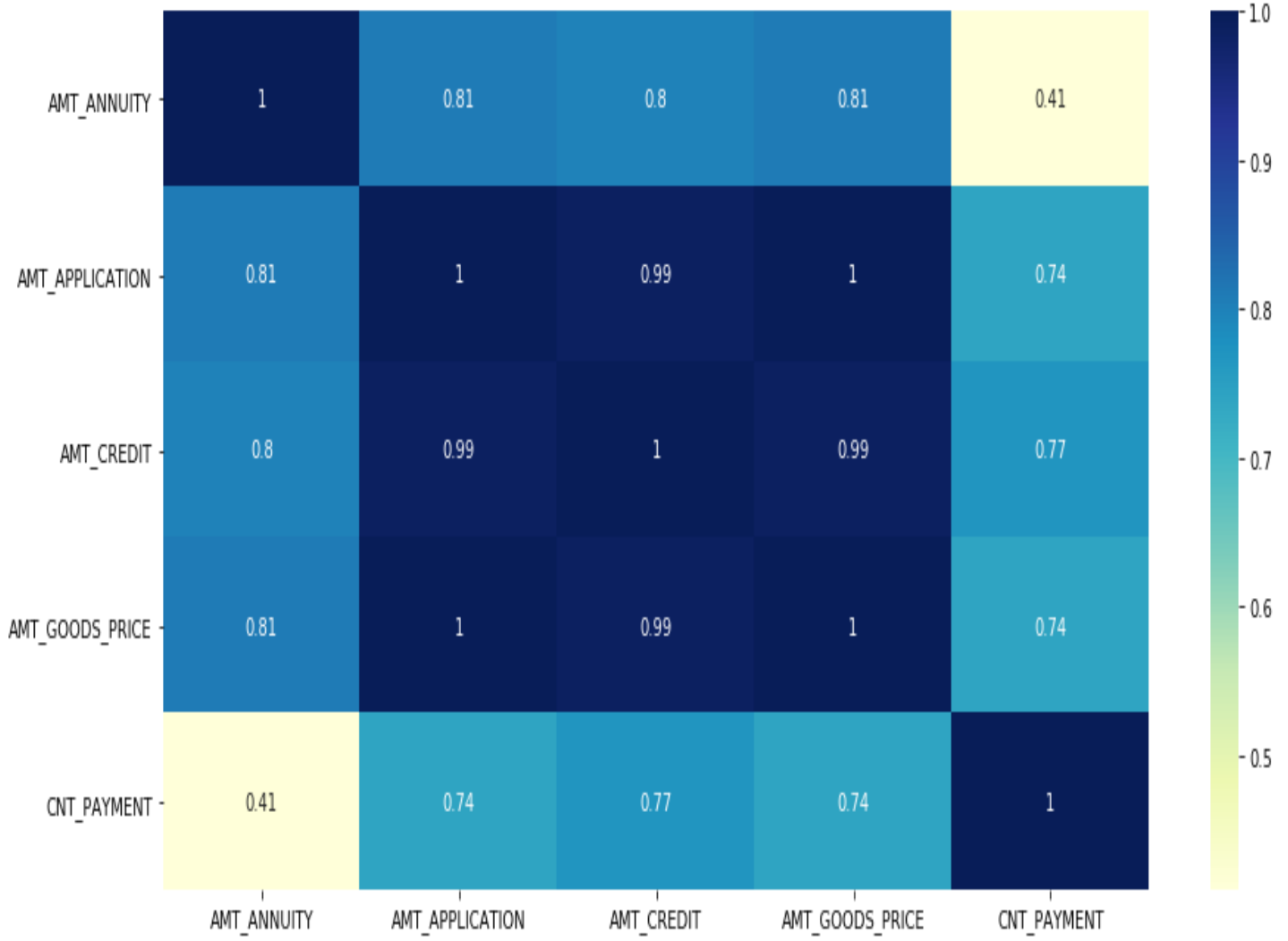






CORRELATION

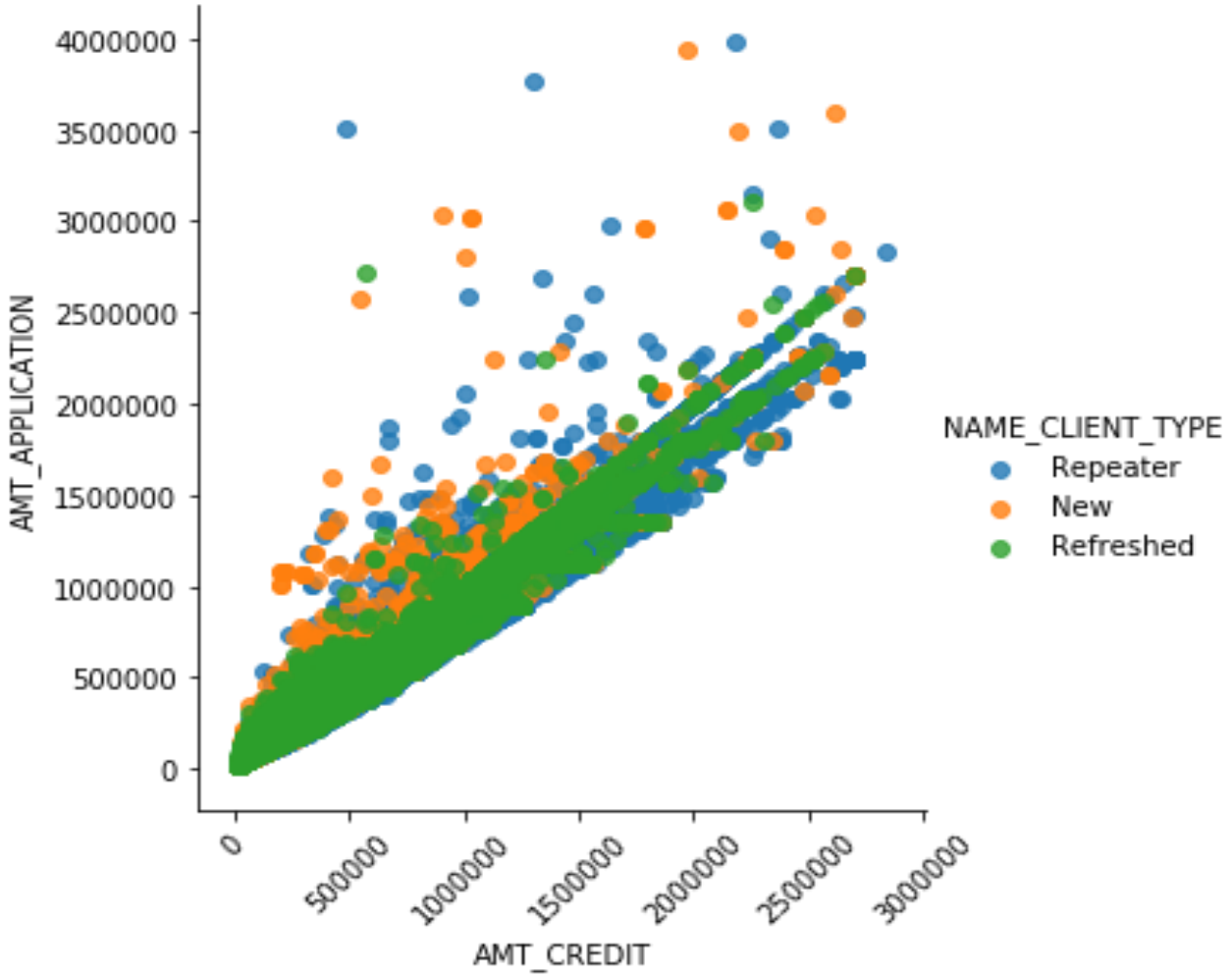
Correlation between credit amount and application amount is very high which translate that the final disbursed amount is very near to the applied amount:



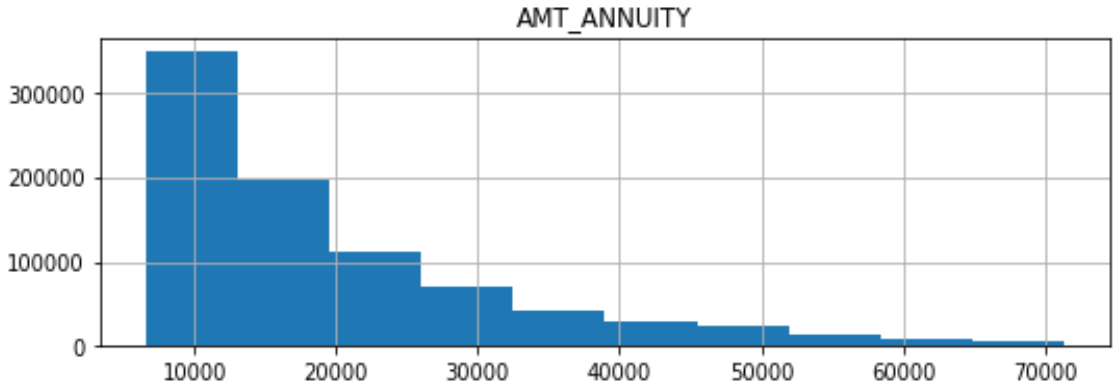
TOP-10 HIGHLY CORRELATED VARIABLE PRESENT IN PREVIOUS-APPLICATION DATASET

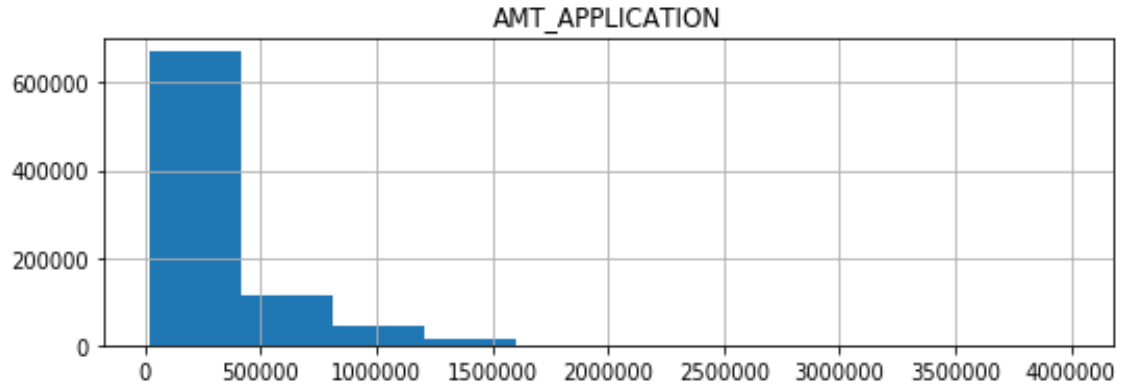
	VAR1	VAR2	CORR
16	AMT_GOODS_PRICE	AMT_APPLICATION	1.000000
11	AMT_CREDIT	AMT_APPLICATION	0.993260
17	AMT_GOODS_PRICE	AMT_CREDIT	0.993260
5	AMT_APPLICATION	AMT_ANNUITY	0.808503
15	AMT_GOODS_PRICE	AMT_ANNUITY	0.808503
10	AMT_CREDIT	AMT_ANNUITY	0.803319
22	CNT_PAYMENT	AMT_CREDIT	0.770284
21	CNT_PAYMENT	AMT_APPLICATION	0.744417
23	CNT_PAYMENT	AMT_GOODS_PRICE	0.744417
20	CNT_PAYMENT	AMT_ANNUITY	0.406836

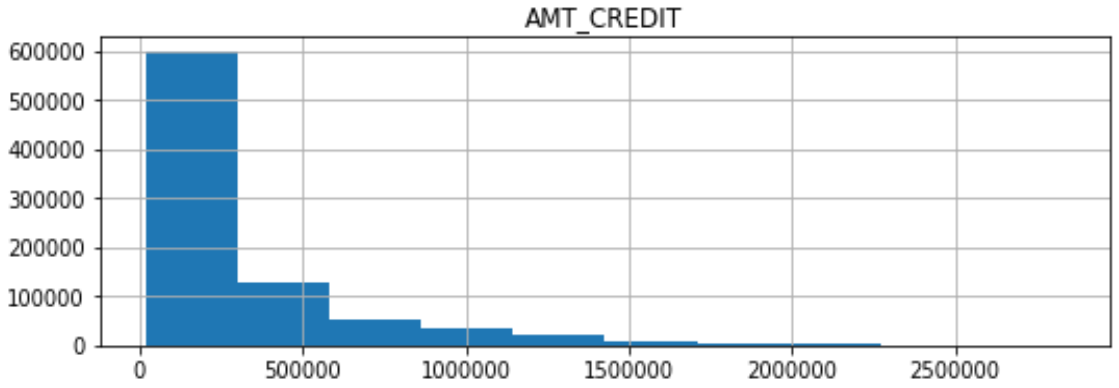
AFTER PLOTING APPLICATION AMOUNT AND CREDIT AMOUNT IT HAS SHOWN US A POSITIVE CORRELATION WHICH MAKE THEM DIRECTLY PROPORTIONAL

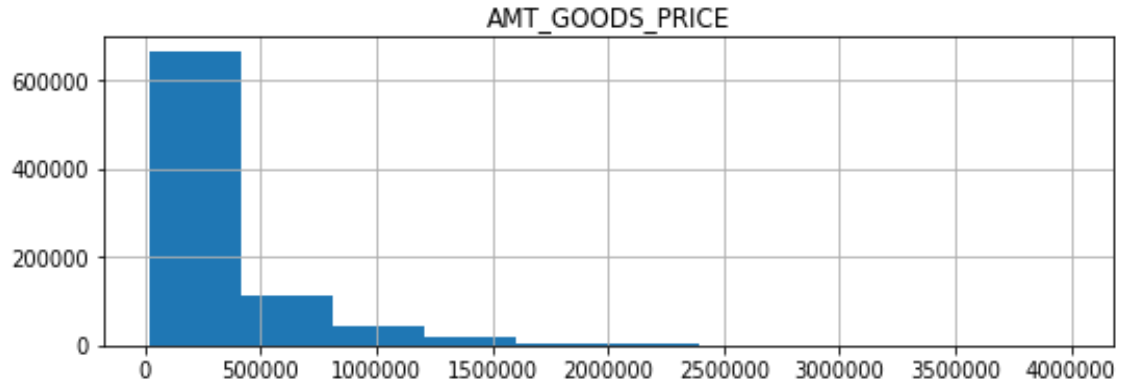


CORRELATION BETWEEN ANNUITY AMOUNT AND APPLICATION AMOUNT IS VERY HIGH WHICH TRANSLATE THAT THE FINAL DISBURSED AMOUNT IS VERY NEAR TO THE APPLIED AMOUNT.









THANK YOU