CREDIT EDA CASE STUDY

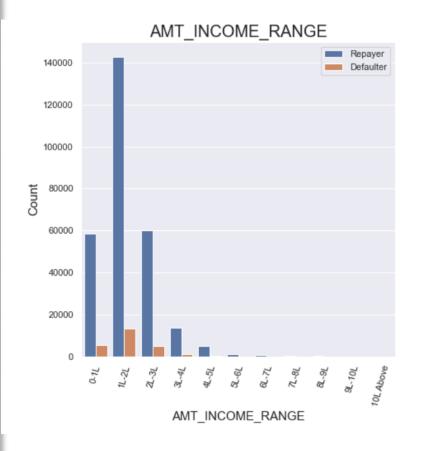
BY SOUMITA PACHHAL

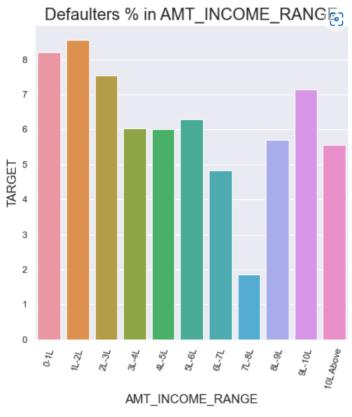
CATEGORICAL UNIVARIATE ANALYSIS

DISTRIBUTION OF INCOME RANGE

Points to be concluded from the graph on the right side.

- Majority of the applications have an Income total of less than 3 Lakhs
- Application with Income less than 3
 Lakhs has a high probability of defaulting
- Applicants with an Income of 7-8
 Lakhs are less likely to default.

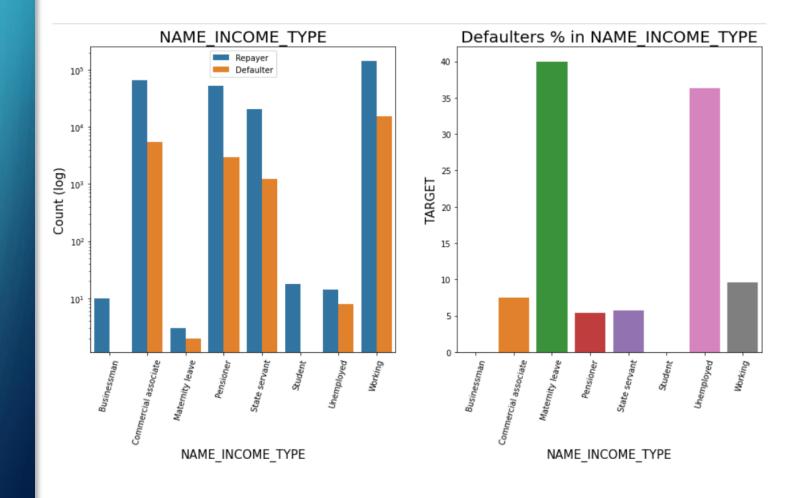




DISTRIBUTION OF INCOME TYPE

Points to be concluded from the graph on the right.

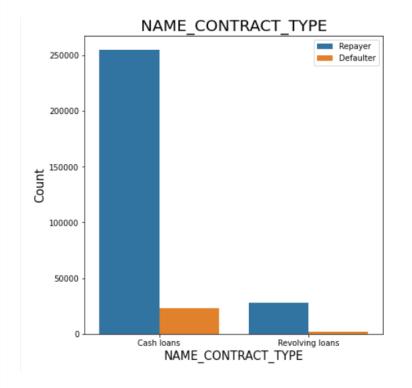
- Most of applicants for loans income type is Working, followed by Commercial associate, Pensioner and State servant.
- The applicants who are on Maternity leave have defaulting percentage of 40% which is the highest, followed by Unemployed (37%). The rest under average around 10% defaultees.
- Student and Businessmen though less in numbers, do not have default record.
 Safest two categories for providing loan

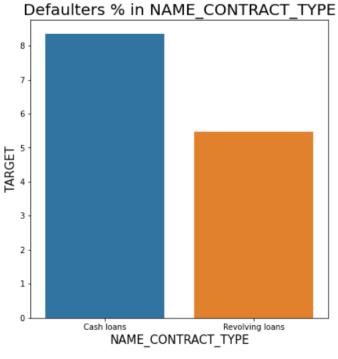


DISTRIBUTION FOR CONTRACT TYPE

Points to be concluded from the graph on the right.

- Revolving loans are just a small fraction (10%) from the total number of loans
- Around 8-9% Cash loan applicants and 5-6% Revolving loan applicant are in defaulters

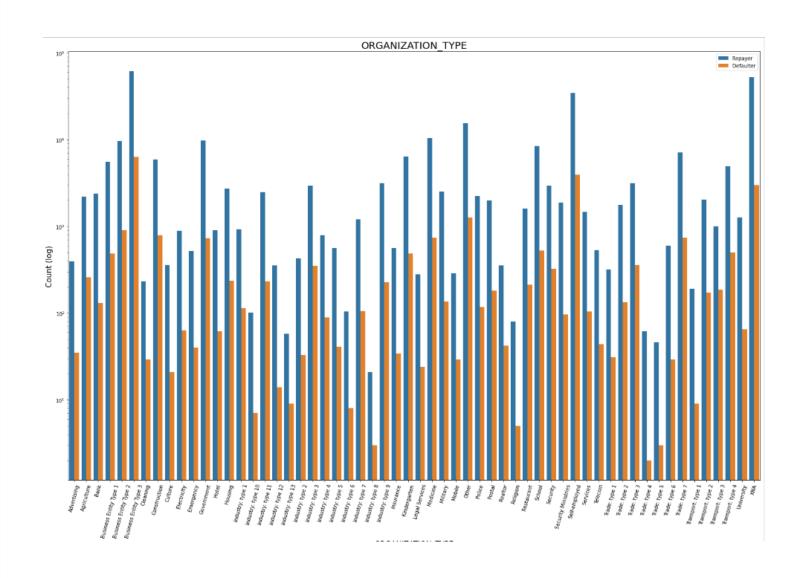




DISTRIBUTION OF ORGANIZATION TYPE

Points to be concluded from the graph on the right.

- Organizations with highest percent of defaultess are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%).
- Self employed people have relative high defaulting rate, to be safer side loan disbursement should be avoided or provide loan with higher interest rate to mitigate the risk of defaulting.
- Most of the people application for loan are from Business Entity Type 3
- For a very high number of applications, Organization type information is unavailable(XNA)
- It can be seen that following category of organization type has lesser defaulters thus safer for providing loans: Trade Type 4 and 5, Industry type 8

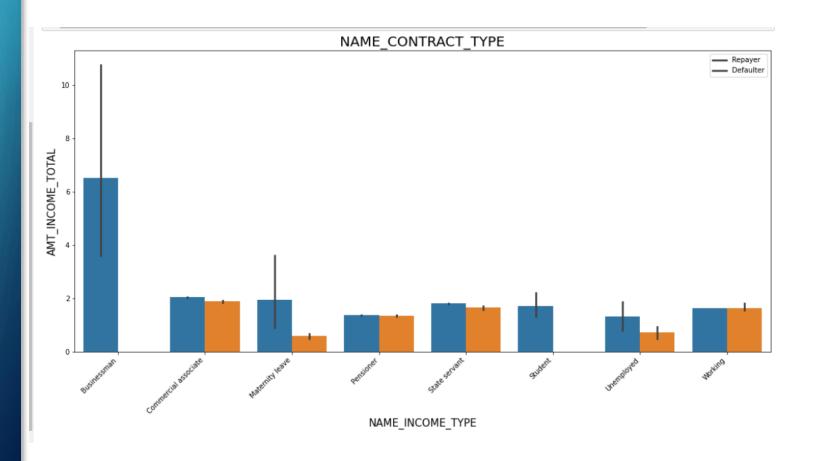


CATEGORICAL BIVARIATE OR MULTIVARIATE ANALYSIS

Income type vs Income Amount Range

Points to be concluded from the graph on the right side.

 It can be seen that Businessman income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a Businessman could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs



Heatmap to see linear correlation among Repayers

Points to be concluded from the graph on the right side.

- 1. Credit amount is highly correlated with:
- Goods Price Amount
- Loan Annuity
- Total Income
- 2. We can also see that repayers have high correlation in number of days employed.

																						_	-1.0
AMT_INCOME_TOTAL -	1	0.34	0.42	0.35	0.17	-0.063	-0.14	-0.065	-0.023	0.077	0.069	-0.028	-0.028	0.041	-0.039	0.0027	0.008	0.0086	0.059	0.018	0.034		
AMT_CREDIT -	0.34	1	0.77	0.99	0.1	0.047	-0.07	-0.013	0.0015	0.054	0.025	0.00089	-0.022	0.07	0.1	-0.0023	0.0051	0.00094	0.055	0.022	-0.038		
AMT_ANNUITY -	0.42	0.77	1	0.78	0.12	-0.012	-0.1	-0.039	-0.014		0.042	-0.013	-0.023		0.1	0.0032		0.012	0.036		-0.008		
AMT_GOODS_PRICE	0.35	0.99	0.78	1	0.1	0.045	-0.069	-0.016	0.0036	0.063	0.027	0.00072	-0.023	0.071	0.079	-0.0017	0.0055	0.0012	0.057	0.022	-0.04		- 0.8
REGION_POPULATION_RELATIVE	0.17	0.1	0.12	0.1	1	0.025	-0.0072	0.052	0.0011	0.17	0.0043	-0.012	0.0023	0.041	-0.086	-0.0023	0.0016	-0.0028	0.071	-0.002	0.00015		
DAYS_BIRTH	-0.063	0.047	-0.012	0.045	0.025	1	0.63	0.33	0.27	-0.096	-0.066	-0.0073	0.00099	0.077	-0.1	-0.0029	-0.0016	0.0036	0.0019	0.015	0.073		
DAYS_EMPLOYED -	-0.14	-0.07	-0.1	-0.069	-0.0072	0.63	1	0.21	0.28	-0.095	-0.038	0.0075	0.016	-0.023	-0.24	-0.0043	0.00093	0.0017	-0.033	0.013	0.047		-0.6
DAYS_REGISTRATION -		-0.013	-0.039	-0.016	0.052	0.33	0.21	1	0.1	0.008	-0.029	-0.0082	-0.0027	0.054	-0.032	0.0025	9.3e-06	0.0013	0.011	0.00036	0.024		
DAYS_ID_PUBLISH -		0.0015	-0.014	0.0036	0.0011	0.27	0.28	0.1	1	-0.034	-0.035	0.013	-0.0025	0.083	-0.05	-0.0019	0.0022	0.0069	0.017	0.017	0.048		
HOUR_APPR_PROCESS_START	0.077	0.054	0.054	0.063	0.17	-0.096	-0.095	0.008	-0.034	1	0.055	-0.008	-0.0088	0.013	-0.013	-0.014	0.0039	-0.0015	0.036	0.0012	-0.025		
REG_REGION_NOT_LIVE_REGION		0.025	0.042	0.027	0.0043	-0.066	-0.038	-0.029	-0.035	0.055	1	-0.02	-0.009	-0.038	-0.034	-0.0016	-0.0012	0.00078	-0.0029	-0.004	-0.018		-0.4
OBS_60_CNT_SOCIAL_CIRCLE	-0.028	28000.0	-0.013	0.00072	-0.012	-0.0073	0.0075	-0.0082	0.013	-0.008	-0.02	1	0.25	0.015	0.027	0.00058	-0.0017	0.001	0.0025	0.0047	0.032		
DEF_60_CNT_SOCIAL_CIRCLE -	-0.028	-0.022	-0.023	-0.023	0.0023	0.00099	0.016	-0.0027	-0.0025	-0.0088	-0.009	0.25	1	0.00016	0.012	-0.002	-0.0016	-0.0022	-0.0014	0.00032	0.016		
DAYS_LAST_PHONE_CHANGE -	0.041	0.07	0.062	0.071	0.041	0.077	-0.023	0.054	0.083	0.013	-0.038	0.015	0.00016	1	0.065	0.0028	0.00081	0.0074	0.045	0.01	0.12		-0.2
FLAG_DOCUMENT_3	-0.039	0.1	0.1	0.079	-0.086	-0.1	-0.24	-0.032	-0.05	-0.013	-0.034	0.027	0.012	0.065	1	0.00025	0.0021	0.0093	0.011	0.011	0.046		
AMT_REQ_CREDIT_BUREAU_HOUR		-0.0023	0.0032	-0.0017	-0.0023	-0.0029	-0.0043	0.0025	-0.0019	-0.014	-0.0016	0.00058	-0.002	0.0028	0.00029	1	0.23	0.0062	0.0034	-3.7e-05	-2.8e-06		
AMT_REQ_CREDIT_BUREAU_DAY		0.0051	0.0025	0.0055	0.0016	-0.0016	0.00093	9.3e-06	0.0022	0.0039	-0.0012	-0.0017	-0.0016	0.00081	0.0021	0.23	1	0.22	-0.0024	-0.002	0.00011		- 0.0
AMT_REQ_CREDIT_BUREAU_WEEK		0.00094	0.012	0.0012	-0.0028	0.0036	0.0017	0.0013	0.0069	-0.0015	0.00078	0.001	-0.0022	0.0074	0.0093	0.0062	0.22	1	-0.0078	-0.0081	0.029		0.0
AMT_REQ_CREDIT_BUREAU_MON -		0.055	0.036	0.057	0.071	0.0019	-0.033	0.011	0.017	0.036	-0.0029	0.0025	-0.0014	0.045	0.011	0.0034	-0.0024	-0.0078	1	0.0045	0.013		
AMT_REQ_CREDIT_BUREAU_QRT	0.018	0.022	0.012	0.022	-0.002	0.015	0.013	0.00036	0.017	0.0012	-0.004	0.0047	-0.00032	0.01	0.011	-3.7e-05	-0.002	-0.0081	0.0045	1	0.093		
AMT_REQ_CREDIT_BUREAU_YEAR		-0.038	-0.008	-0.04	0.00015	0.073	0.047	0.024	0.048	-0.025	-0.018	0.032	0.016	0.12	0.046	-2.8e-06	0.00011	0.029	0.013	0.093	1		0.2
	AMT_INCOME_TOTAL -	AMT_CREDIT -	AMT_ANNUITY -	AMT_GOODS_PRICE -	I_POPULATION_RELATIVE -	DAYS_BIRTH -	DAYS_EMPLOYED -	DAYS_REGISTRATION -	DAYS ID PUBLISH -	R_APPR_PROCESS_START -	GION NOT LIVE REGION -	60_CNT_SOCIAL_CIRCLE -	60 CNT SOCIAL CIRCLE -	S_LAST_PHONE_CHANGE -	FLAG_DOCUMENT_3 -	CREDIT_BUREAU_HOUR -	:Q_CREDIT_BUREAU_DAY -	CREDIT_BUREAU_WEEK -	2_CREDIT_BUREAU_MON -	:Q_CREDIT_BUREAU_QRT -)_CREDIT_BUREAU_YEAR -		



AMT_INCOME_TOTAL -	1	0.038	0.046	0.038	0.0091	-0.0031	-0.015	0.00016	0.0042	0.014	0.0076	-0.0046	-0.0049	0.0024	0.0014	0.0011	0.00014	0.00094	0.0057	0.001	0.0045		-1.0
AMT_CREDIT -	0.038	1	0.75	0.98	0.069	0.14	0.0019	0.026	0.052	0.032	0.02	0.019	-0.031	0.11	0.062	-0.0038	0.0043	0.011	0.056	-0.0072	-0.021		
AMT_ANNUITY -	0.046	0.75	1	0.75	0.072	0.014	-0.081	-0.034	0.017	0.031	0.035	0.0055	-0.027	0.08	0.1	0.013	7.4e-05	0.029	0.049	-0.0073	-0.0098		
AMT_GOODS_PRICE -	0.038	0.98	0.75	1	0.076	0.14	0.0066	0.026	0.056	0.044	0.022	0.02	-0.026	0.12	0.038	-0.0025	0.0054	0.011	0.059	-0.0061	-0.023		- 0.8
REGION_POPULATION_RELATIVE	0.0091	0.069	0.072	0.076	1	0.048	0.016	0.056	0.016	0.14	-0.022	0.0068	0.018	0.055	-0.042	0.00071	-0.0045	0.0035	0.065	-0.0044	0.0032		
DAYS_BIRTH -	-0.0031	0.14	0.014	0.14	0.048	1	0.58	0.29	0.25	-0.062	-0.055	0.0054	-0.004	0.11	-0.13	-0.012	0.008	0.0082	0.011	0.023	0.084		
DAYS_EMPLOYED -	-0.015	0.0019	-0.081	0.0066	0.016	0.58	1	0.19	0.23	-0.06	-0.035	-0.0083	0.0054	0.0014	-0.27	-0.0073	0.019	0.013	-0.023	0.012	0.031		- 0.6
DAYS_REGISTRATION -	0.0001€	0.026	-0.034	0.026	0.056	0.29	0.19	1	0.097	0.033	-0.02	-0.015	-0.0092	0.072	-0.038	0.0048	0.0079	-0.0023	-0.0034	0.013	0.014		
DAYS_ID_PUBLISH -	0.0042	0.052	0.017	0.056	0.016	0.25	0.23	0.097	1	-0.022	-0.033	0.02	-0.0049	0.12	-0.027	0.00048	0.014	0.0037	0.024	0.008	0.056		
HOUR_APPR_PROCESS_START -	0.014	0.032	0.031	0.044	0.14	-0.062	-0.06	0.033	-0.022	1	0.052	-0.012	-0.0003	0.023	-0.008	-0.015	-0.014	-0.0041	0.039	0.0032	-0.023		-0.4
REG_REGION_NOT_LIVE_REGION -	0.0076	0.02	0.035	0.022	-0.022	-0.055	-0.035	-0.02	-0.033	0.052	1	-0.028	0.0016	-0.034	-0.033	-0.0095	-0.006	0.0075	0.019	0.00099	-0.023		0.4
OBS_60_CNT_SOCIAL_CIRCLE -	-0.0046	0.019	0.0055	0.02	0.0068	0.0054	-0.0083	-0.015	0.02	-0.012	-0.028	1	0.26	0.029	0.015	0.0018	-0.0081	-0.0017	0.0061	0.0087	0.038		
DEF_60_CNT_SOCIAL_CIRCLE -	-0.0049	-0.031	-0.027	-0.026	0.018	-0.004	0.0054	-0.0092	-0.0049	-0.0003	0.0016	0.26	1	-0.0046	-0.009	0.0039	-0.0037	-0.0051	0.00027	-0.0023	0.00091		
DAYS_LAST_PHONE_CHANGE -	0.0024	0.11	0.08	0.12	0.055	0.11	0.0014	0.072	0.12	0.023	-0.034	0.029	-0.0046	1	0.051	0.0024	0.0044	0.0094	0.05	0.003	0.11		-0.2
FLAG_DOCUMENT_3 -	0.0014	0.062	0.1	0.038	-0.042	-0.13	-0.27	-0.038	-0.027	-0.008	-0.033	0.015	-0.009	0.051	1	-0.002	-0.012	-0.0035	0.02	0.0061	0.041		
AMT_REQ_CREDIT_BUREAU_HOUR -	0.0011	-0.0038	0.013	-0.0025	0.00071	-0.012	-0.0073	0.0048	0.00048	-0.015	-0.0095	0.0018	0.0039	0.0024	-0.002	1	0.25	0.0085	-0.0037	0.012	0.0045		
AMT_REQ_CREDIT_BUREAU_DAY	0.00014	0.0043	7.4e-05	0.0054	-0.0045	0.008	0.019	0.0079	0.014	-0.014	-0.006	-0.0081	-0.0037	0.0044	-0.012	0.25	1	0.19	-0.0086	0.006	0.008		- 0.0
AMT_REQ_CREDIT_BUREAU_WEEK	0.00094	0.011	0.029	0.011	0.0035	0.0082	0.013	-0.0023	0.0037	-0.0041	-0.0075	-0.0017	-0.0051	0.0094	-0.0035	0.0085	0.19	1	-0.0029	0.0018	0.032		
AMT_REQ_CREDIT_BUREAU_MON -	0.0057	0.056	0.049	0.059	0.065	0.011	-0.023	-0.0034	0.024	0.039	0.019	0.0061	0.00027	0.05	0.02	-0.0037	-0.0086	-0.0029	1	0.02	0.025		
AMT_REQ_CREDIT_BUREAU_QRT -	0.001	-0.0072	-0.0073	-0.0061	-0.0044	0.023	0.012	0.013	0.008	0.0032	0.00099	0.0087	-0.0023	0.003	0.0061	0.012	0.006	0.0018	0.02	1	0.13		0.2
AMT_REQ_CREDIT_BUREAU_YEAR -	0.0045	-0.021	-0.0098	-0.023	0.0032	0.084	0.031	0.014	0.056	-0.023	-0.023	0.038	0.00091	0.11	0.041	0.0045	0.008	0.032	0.025	0.13	1		
	AMT_INCOME_TOTAL -	AMT_CREDIT -	AMT_ANNUITY -	AMT_GOODS_PRICE -	V_POPULATION_RELATIVE -	DAYS_BIRTH -	DAYS_EMPLOYED -	DAYS_REGISTRATION -	DAYS_ID_PUBLISH -	R APPR PROCESS START -	GION_NOT_LIVE_REGION -	60 CNT SOCIAL CIRCLE	60 CNT SOCIAL CIRCLE -	S_LAST_PHONE_CHANGE -	FLAG_DOCUMENT_3 -	CREDIT_BUREAU_HOUR -	EO_CREDIT_BUREAU_DAY .	CREDIT BUREAU WEEK -	Q_CREDIT_BUREAU_MON -	SQ_CREDIT_BUREAU_QRT -	2 CREDIT BUREAU YEAR -		

CORRELATION

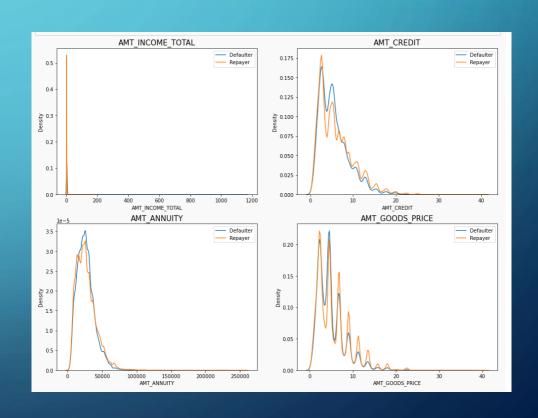
Points to be concluded from the graph presented before.

- Credit amount is highly correlated with good price amount which is same as repayers.
- Loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayers(0.77)
- We can also see that repayers have high correlation in number of days employed (0.62) when compared to defaulters (0.58).
- There is a severe drop in the correlation between total income of the client and the credit amount (0.038) amongst defaulters whereas it is 0.342 among repayers.
- Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in repayers.
- There is a slight increase in defaulted to observed count in social circle among defaulters(0.264) when compared to repayers(0.254)

NUMERICAL UNIVARIATE ANALYSIS

DISTRIBUTION PLOT

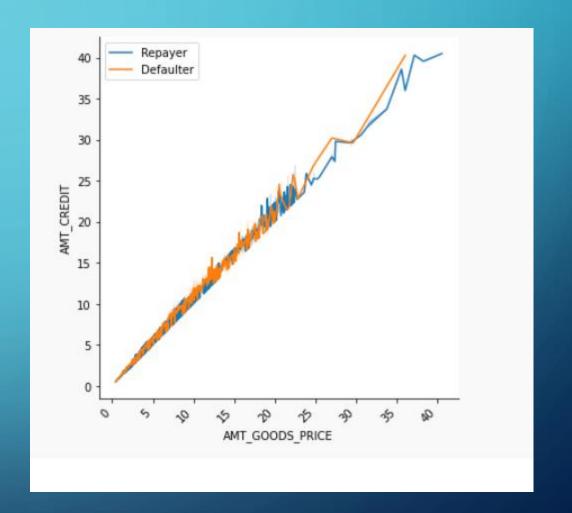
- Most no of loans are given for goods price below 10 lakhs
- Most people pay annuity below 50K for the credit loan
- Credit amount of the loan is mostly less then 10 lakhs
- The repayers and defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision



NUMERICAL BIVARIATE ANALYSIS

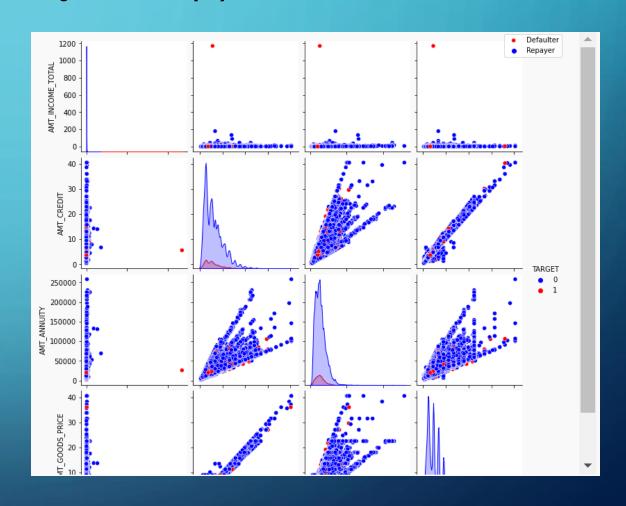
Few points can be concluded from the graph.

 When the credit amount goes beyond 30 Lakhs, there is an increase in defaulters



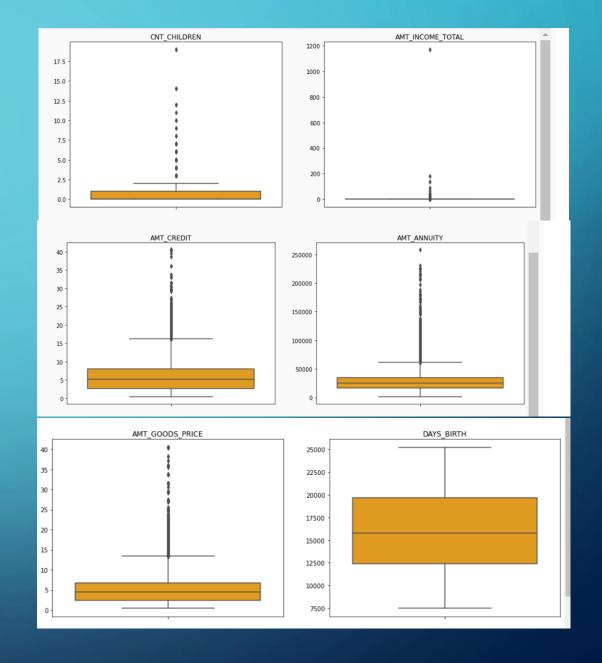
Pairplot between amount variable to draw reference against loan repayment status

- When Annuity Amount > 15K and Good Price Amount > 20 Lakhs, there is a lesser chance of defaulters
- Loan Amount(AMT_CREDIT) and Goods
 price(AMT_GOODS_PRICE) are highly
 correlated as based on the scatterplot where
 most of the data are consolidated in form of a
 line
- There are very less defaulters for AMT_CREDIT
 20 Lakhs



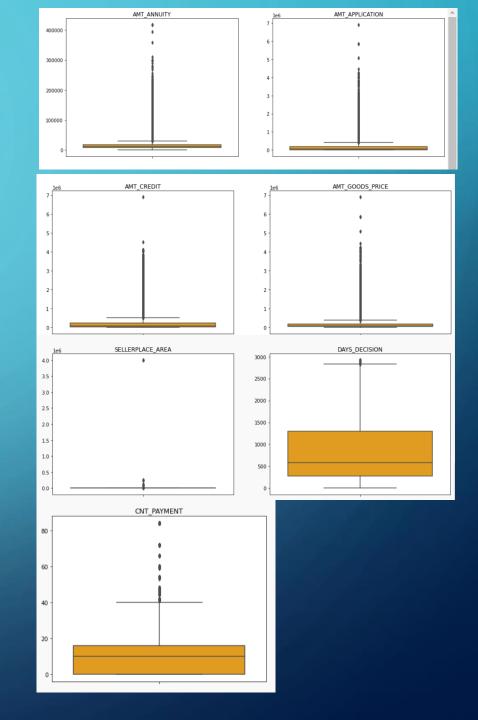
BOXPLOT FOR APPLICATION_DATA

- AMT_ANNUITY, AMT_CREDIT,
 AMT_GOODS_PRICE,CNT_CHILDREN have some
 number of outliers.
- AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have high income when compared to the others.
- DAYS_BIRTH has no outliers which means the data available is reliable.
- DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.



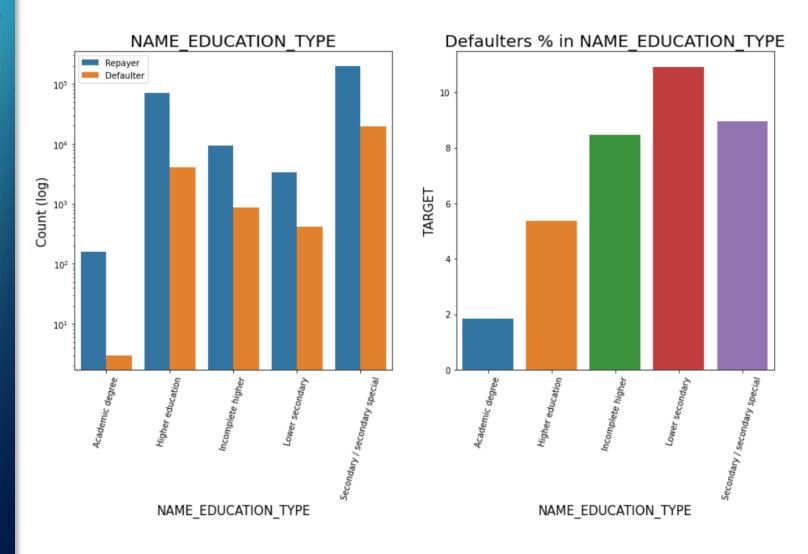
BOXPLOT FOR PREVIOUS_APPLICATION

- AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.
- CNT_PAYMENT has few outlier values.
- DAYS_DECISION has little number of outliers indicating that these previous applications decisions were taken long back.



Analyzing Education Type based on loan repayment status

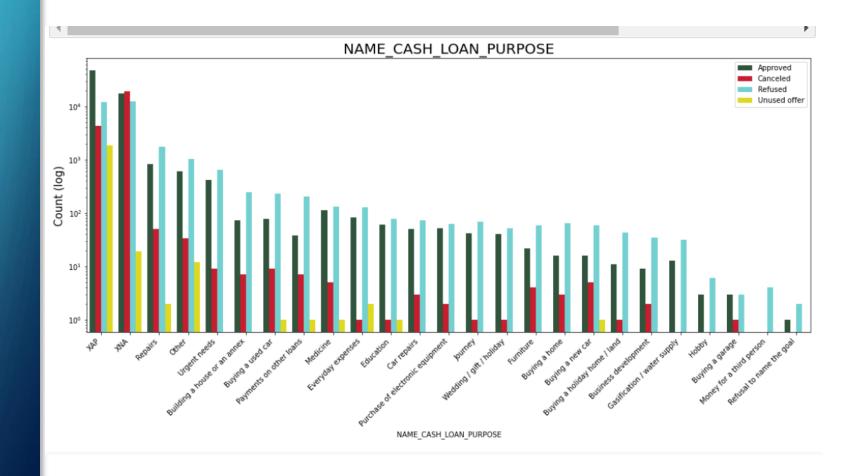
- Majority of clients have Secondary/secondary special education, followed by clients with Higher education.
- Very few clients have an academic degree
- Lower secondary category have highest rate of defaulting around 11%.
- People with Academic degree are least likely to default.





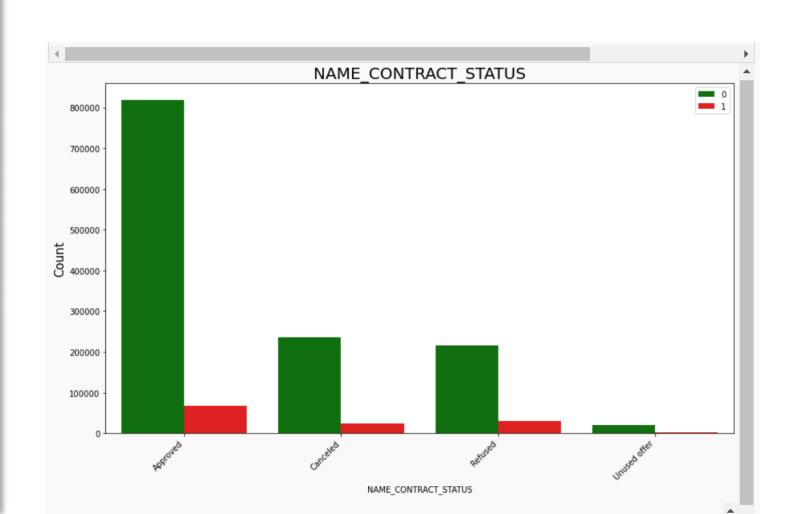
CONTRACT STATUS VS PURPOSE OF THE LOAN

- Loan purpose has high number of unknown values (XAP, XNA)
- Loan taken for the purpose of Repairs looks to have highest default rate
- rejected by bank or refused by client which are applied for Repair or Other. from this we can infer that repair is considered high risk by bank. Also, either they are rejected or bank offers loan on high interest rate which is not feasible by the clients and they refuse the loan



Contract Status based on loan repayment status

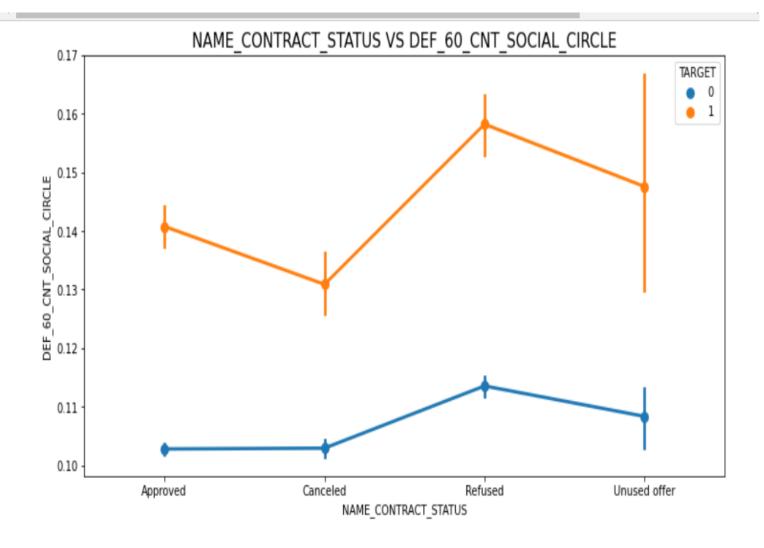
- 90% of the previously cancelled client have actually repayed the loan. Revising the interest rates would increase business opportunity for these clients
- 88% of the clients who have been previously refused a loan has payed back the loan in current case.
- Refusal reason should be recorded for further analysis as these clients could turn into potential repaying customer.



Relationship between people who defaulted in the last 60 days being in the client's social circle and contact status

Few points can be concluded from the graph.

 Clients who have average of 0.13 or higher their
 DEF_60_CNT_SOCIAL_CIRCLE score tend to default more and thus analysing client's social circle could help in the disbursement of the loan



CONCLUSION

Decisive Factor whether an applicant will be Repayer:

- NAME_EDUCATION_TYPE: Academic degree has less defaults.
- NAME_INCOME_TYPE: Student and Businessmen have no defaults.
- REGION_RATING_CLIENT: RATING 1 is safer.
- ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%
- DAYS_BIRTH: People above age of 50 have low probability of defaulting
- DAYS_EMPLOYED: Clients with 40+ year experience having less than 1% default rate
- AMT_INCOME_TOTAL:Applicant with Income more than 700,000 are less likely to default
- NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, Buying garage are being repayed mostly.
- CNT_CHILDREN: People with zero to two children tend to repay the loans

Decisive Factor whether an applicant will be Defaulter

- CODE_GENDER: Men are at relatively higher default rate
- NAME_FAMILY_STATUS: People who have civil marriage or who are single default a lot.
- NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education
- NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot.
- REGION_RATING_CLIENT: People who live in Rating 3 has highest defaults.
- OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as their default rate is huge.
- ORGANIZATION_TYPE: Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed people have relative high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.
- DAYS_BIRTH: Avoid young people who are in age group of 20-40 as they have higher probability of defaulting
- DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.
- CNT_CHILDREN & CNT_FAM_MEMBERS: Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected.
- AMT_GOODS_PRICE: When the credit amount goes beyond 3lakhs, there is an increase in defaulters.

Factors that Loan can be given on Condition of High-Interest rate to mitigate any default risk leading to business loss:

- NAME_HOUSING_TYPE: High number of loan applications are from the category of people who live in Rented apartments & living with parents and hence offering the loan would mitigate the loss if any of those default.
- AMT_CREDIT: People who get loan for 3-6 Lakhs tend to default more than others and hence having higher interest specifically for this credit range would be ideal.
- AMT_INCOME: Since 90% of the applications have Income total less than 3Lakhs and they have high probability of defaulting, they could be offered loan with higher interest compared to other income category.
- CNT_CHILDREN & CNT_FAM_MEMBERS: Clients who have 4 to 8 children has a very high default rate and hence higher interest should be imposed on their loans.
- NAME_CASH_LOAN_PURPOSE: Loan taken for the purpose of Repairs seems to have highest default rate. A very high number of applications have been rejected by bank or refused by client in previous applications as well which has a purpose as repair or other. This shows that purpose repair is taken as high risk by bank and either they are rejected, or bank offers very high loan interest rate which is not feasible by the clients, thus they refuse the loan. The same approach could be followed in future as well.

• 90% of the previously canceled client have actually repaid the loan. Record the reason for cancellation which might help the bank to determine and negotiate terms with these repaying customers in the future to increase business opportunities. 88% of the clients who were refused by banks for loans earlier have now turned into repaying clients. Hence documenting the reason for rejection could mitigate the business loss and these clients could be contacted for further loans.

