EDA: Bank Default Risk Analysis

Overview

The increasing trends in loan repayment default have been a matter of grave concern for policy makers and banks at large over a long period of time, defaulting on loan repayment not only affects safety of deposits, profitability of banks but the overall financial stability of a country as well. We aim to show how EDA can prove useful in the area of risk analytics in the banking and financial services sector by helping to minimize the risk of financial loss caused due to defaults on loan repayment.

Business Understanding

Losing business due to loan refusal

Losing business due to loan grant

Risks in Providing Loans

Repayment of Loans

- Payment difficulties--Late Payments
- Payment on Time

Possible Decisions

- Approved
- Cancelled
- Refused
- Unused Offer

Business Objective

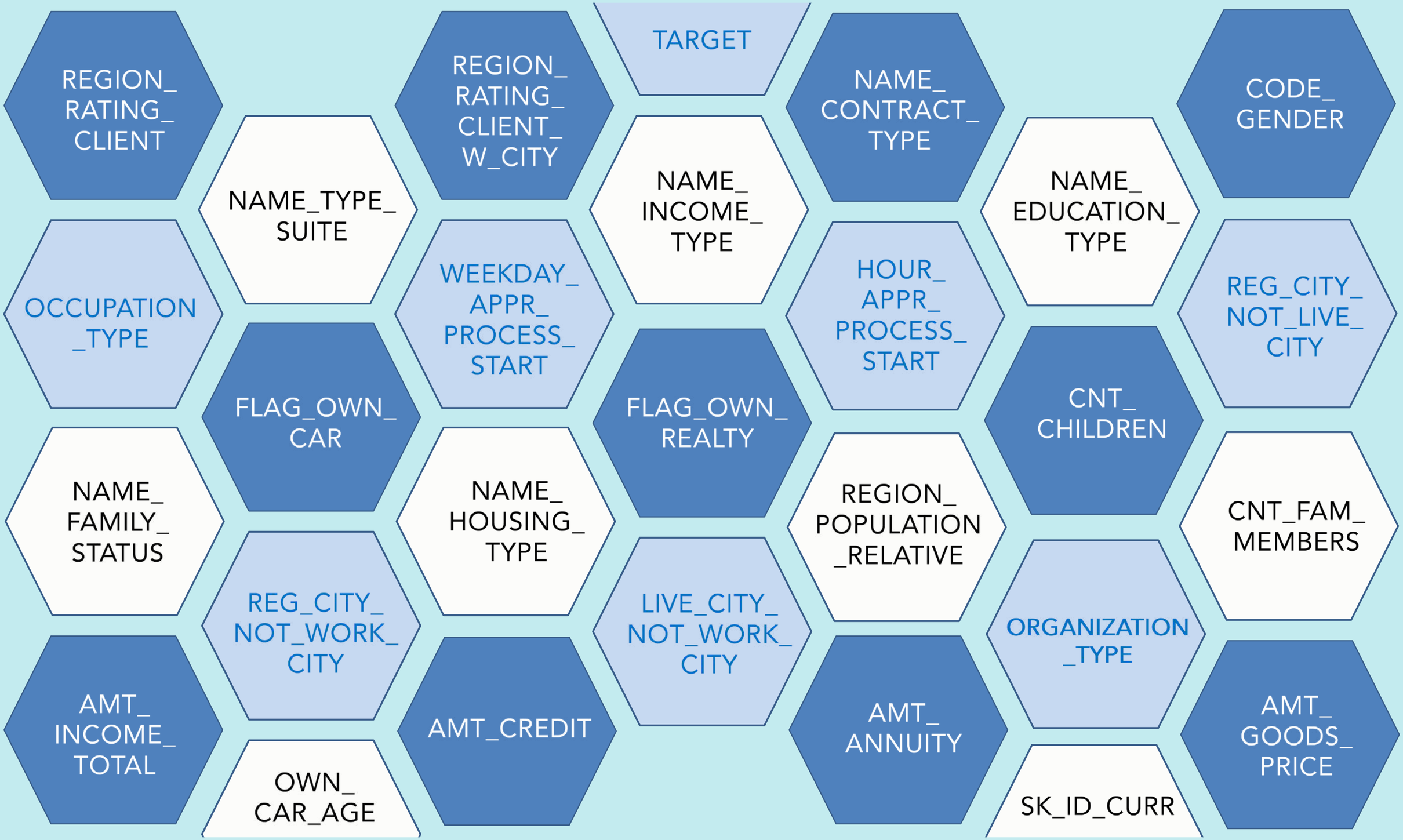
Primary Objective

We aim to identify the pattern of installment payment by client so that to make decisions such as loan approval, denying, reducing the amount of the loan, leading at certain interest based on the basis of risk associated with it. This will also ensure that only borrowers who can repay the loan will be accepted.

Secondary Objective

Identifying driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

DATA DESCRIPTION



Exploratory Data Analytics

Commercial associate

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
mean	188217.0	607288.0	28893.0	531910.0
median	157500.0	526491.0	27189.0	450000.0
std	108316.0	384218.0	13389.0	345861.0
min	36000.0	45000.0	4540.0	45000.0
max	1890000.0	4027680.0	106380.0	3600000.0

Maternity leave

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
mean	58500.0	929250.0	26091.0	929250.0
median	58500.0	929250.0	26091.0	929250.0
std	12728.0	715946.0	20937.0	715946.0
min	49500.0	423000.0	11286.0	423000.0
max	67500.0	1435500.0	40896.0	1435500.0

Pensioner

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
mean	135557.0	558039.0	23729.0	492915.0
median	121500.0	500211.0	22185.0	450000.0
std	73041.0	349482.0	11916.0	318022.0
min	25650.0	45000.0	2722.0	45000.0
max	1260000.0	2173500.0	149211.0	2173500.0

State servant

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
mean	164713.0	614816.0	27546.0	541962.0
median	148500.0	540000.0	26217.0	450000.0
std	117514.0	370360.0	12581.0	337375.0
min	27900.0	45000.0	4320.0	45000.0
max	3150000.0	2013840.0	103455.0	1980000.0

Unemployed

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
mean	72000.0	626625.0	22626.0	626625.0
median	65250.0	578250.0	22270.0	578250.0
std	34271.0	276644.0	8469.0	276644.0
min	31500.0	328500.0	10629.0	328500.0
max	135000.0	1215000.0	38840.0	1215000.0

Working

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
mean	163677.0	535532.0	26087.0	468603.0
median	135000.0	469152.0	25034.0	450000.0
std	949989.0	326900.0	12044.0	292776.0
min	27000.0	45000.0	2844.0	45000.0
max	117000000.0	2695500.0	127508.0	2254500.0

We have grouped our data to understand the income type groups which have higher amount defaults in loan repayment. We could find "Commercial Associate" to be the income group with highest mean total income. However, the mean Amount of Credit is of "Maternity Leave" Income Groups, followed by "Unemployed" and "Commercial Associate".

Exploratory Data Analytics

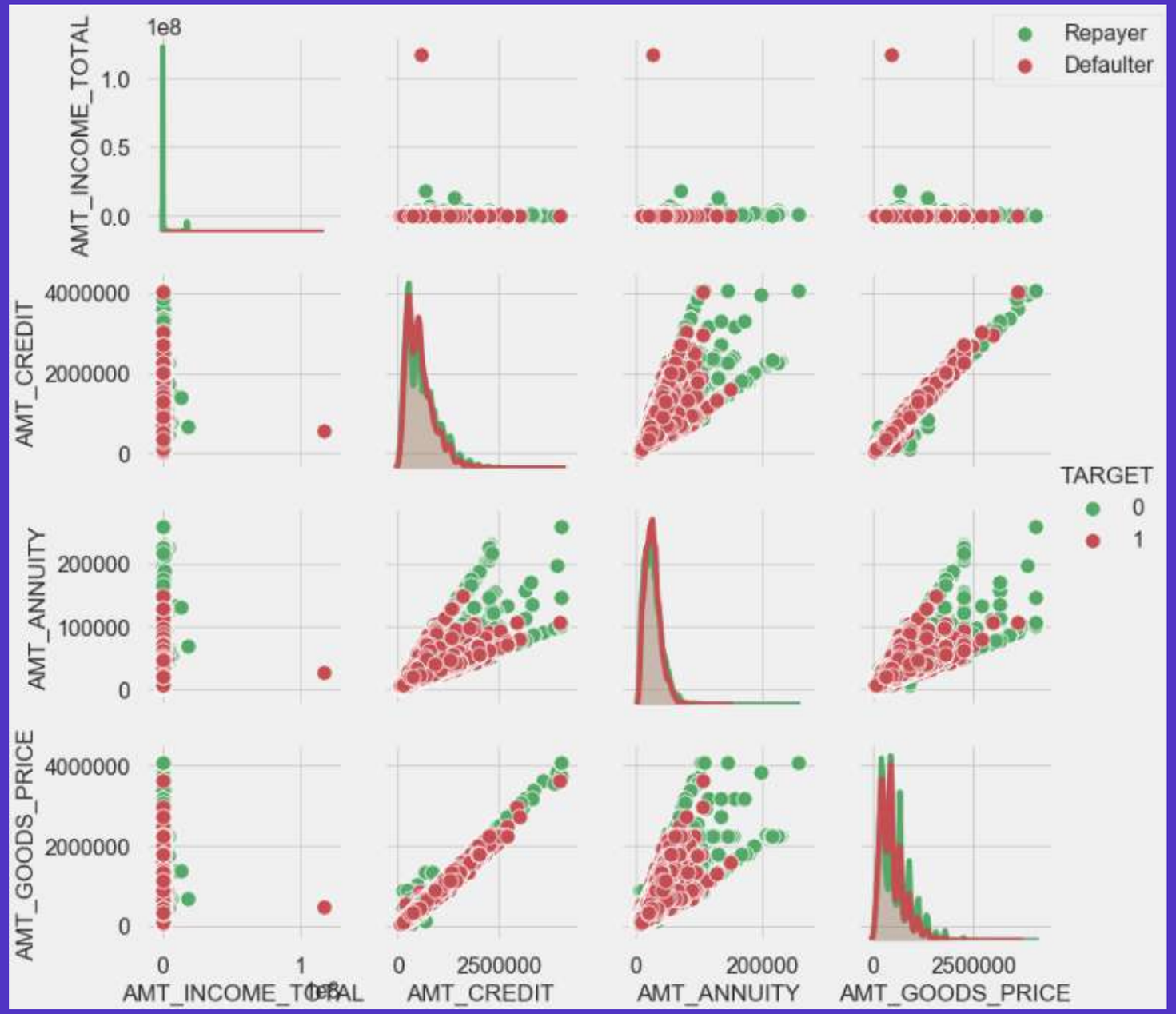
NAME_INCOME_TYPE	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
Commercial associate	188217.0	607288.0	28893.0	531910.0
Pensioner	135557.0	558039.0	23729.0	492915.0
State servant	164713.0	614816.0	27546.0	541962.0
Working	163677.0	535532.0	26087.0	468603.0

Since "Unemployed" and "Maternity Leave" are two income group types which are currently not working, we wanted to focus more on the income groups who have defaults inspite of being in the working class. By the mean data, we understood that "State Servant" Group has the highest mean for the Loan Amount. We could also observe that the Loan Amount credited to all the income groups were about 3-4 times their total incomes. This could also be one of the reasons for the increasing defaults.

```
Skewness :
 AMT_INCOME_TOTAL     154.3468
AMT_CREDIT              1.3339
AMT_ANNUITY             1.0173
AMT_GOODS_PRICE         1.4850
dtype: float64
Kurtosis :
 AMT_INCOME_TOTAL    24150.7878
AMT_CREDIT              2.7122
AMT_ANNUITY             2.4059
AMT_GOODS_PRICE         3.3609
dtype: float64
```

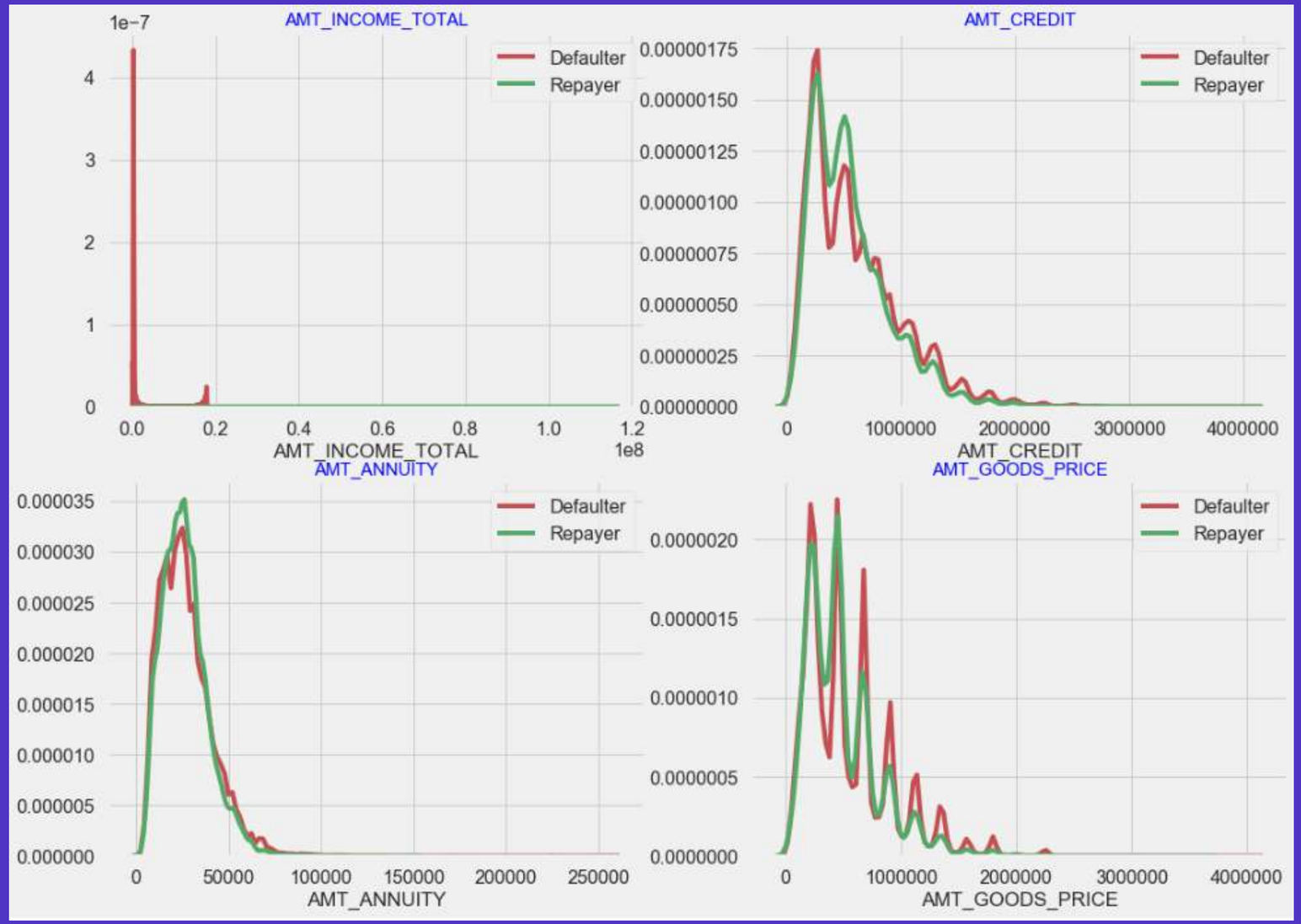
Since "Unemployed" and "Maternity Leave" are two income group types which are currently not working, we wanted to focus more on the income groups who have defaults inspite of being in the working class. By the mean data, we understood that "State Servant" Group has the highest mean for the Loan Amount. We could also observe that the Loan Amount credited to all the income groups were about 3-4 times their total incomes. This could also be one of the reasons for the increasing defaults.

DATA VISUALIZATION



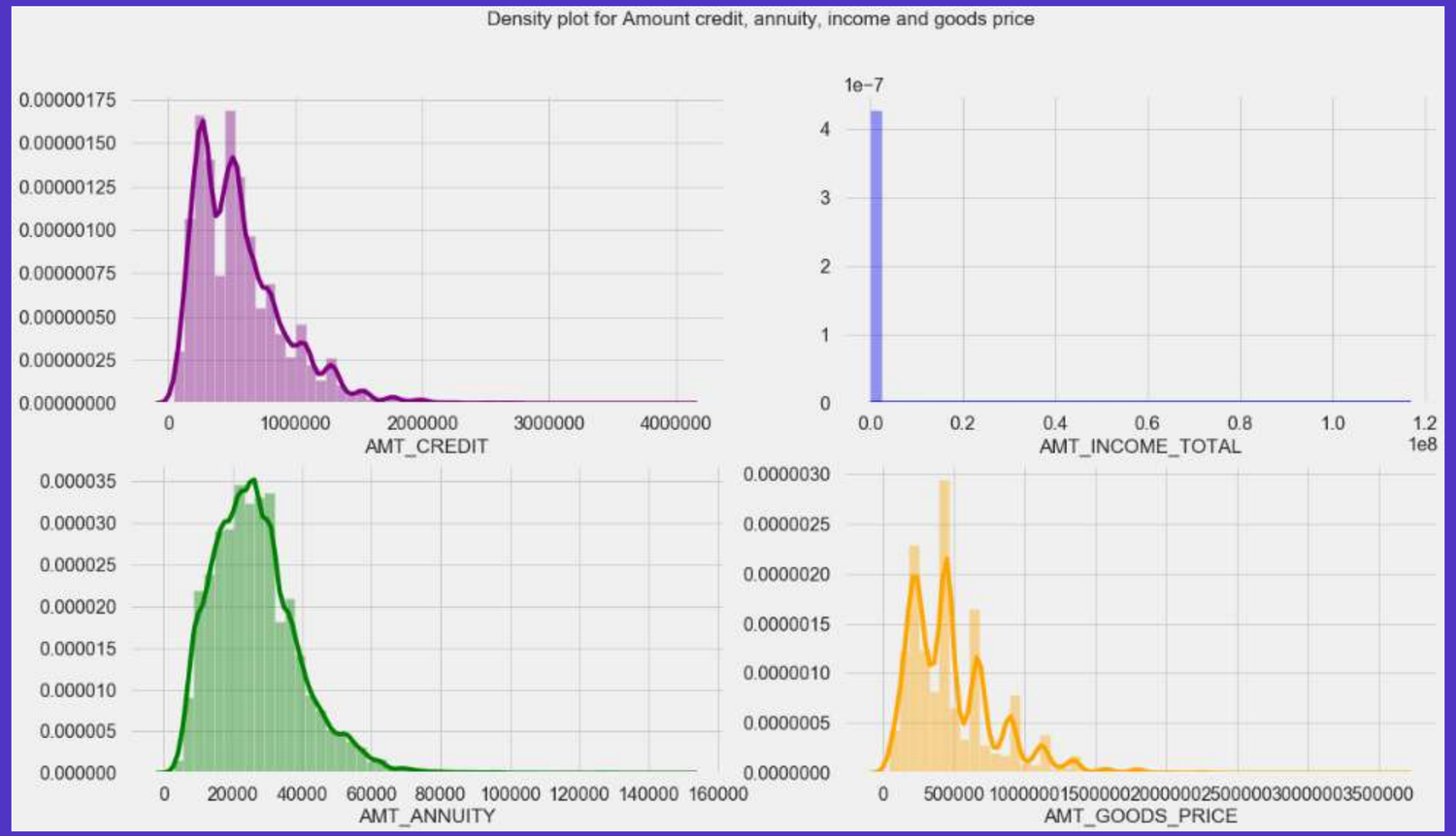
When AMT_ANNUITY >15000 AMT_GOODS_PRICE> 3M, there are a lesser chance of our clients defaulting on the loan repayment. The Credit Amount and the Price of Goods are highly correlated as based on the scatterplot, where most of the data are consolidated in form of a line. There are very less defaulters when the Credit amount is greater than 3M.

DATA VISUALIZATION



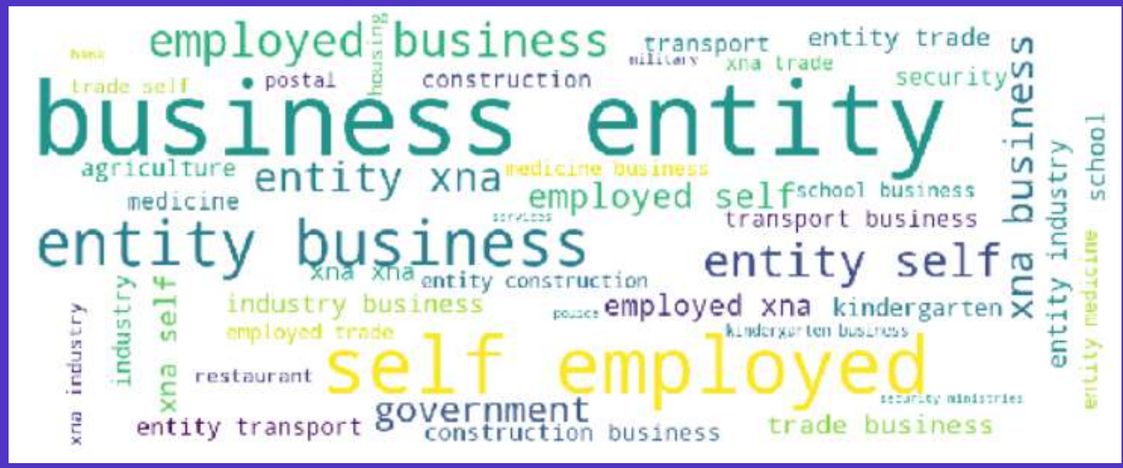
The Distribution plots give a very similar inference as the Pair Plots. We could find higher concentration of defaulters income to be in lower values, higher number of defaults have been made with Amount of Credit lesser than the median, and there is a high concentration and variation in the Price of goods on which the loan had been sanctioned.

DATA VISUALIZATION



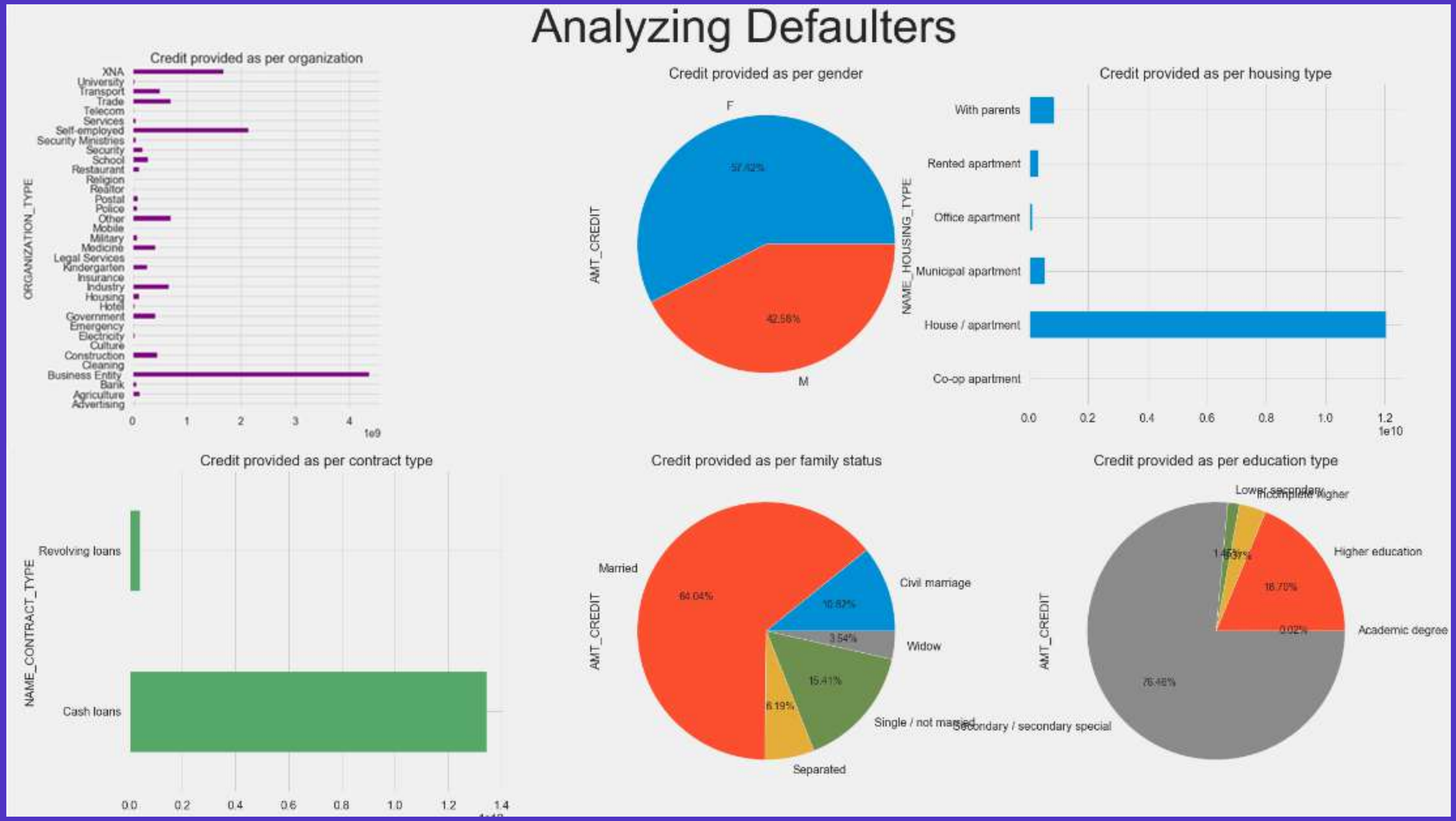
The above plot is a Density Plot of the same distribution plot inferred earlier. The plot helps us understand the density of each measure type.

We have used the Word Cloud to have a visual understanding on the organization types which are prominent in the defaulters. The Word Cloud helped use understand that Self Employed and Business Entity are the most prominent organization types in the defaulters list.



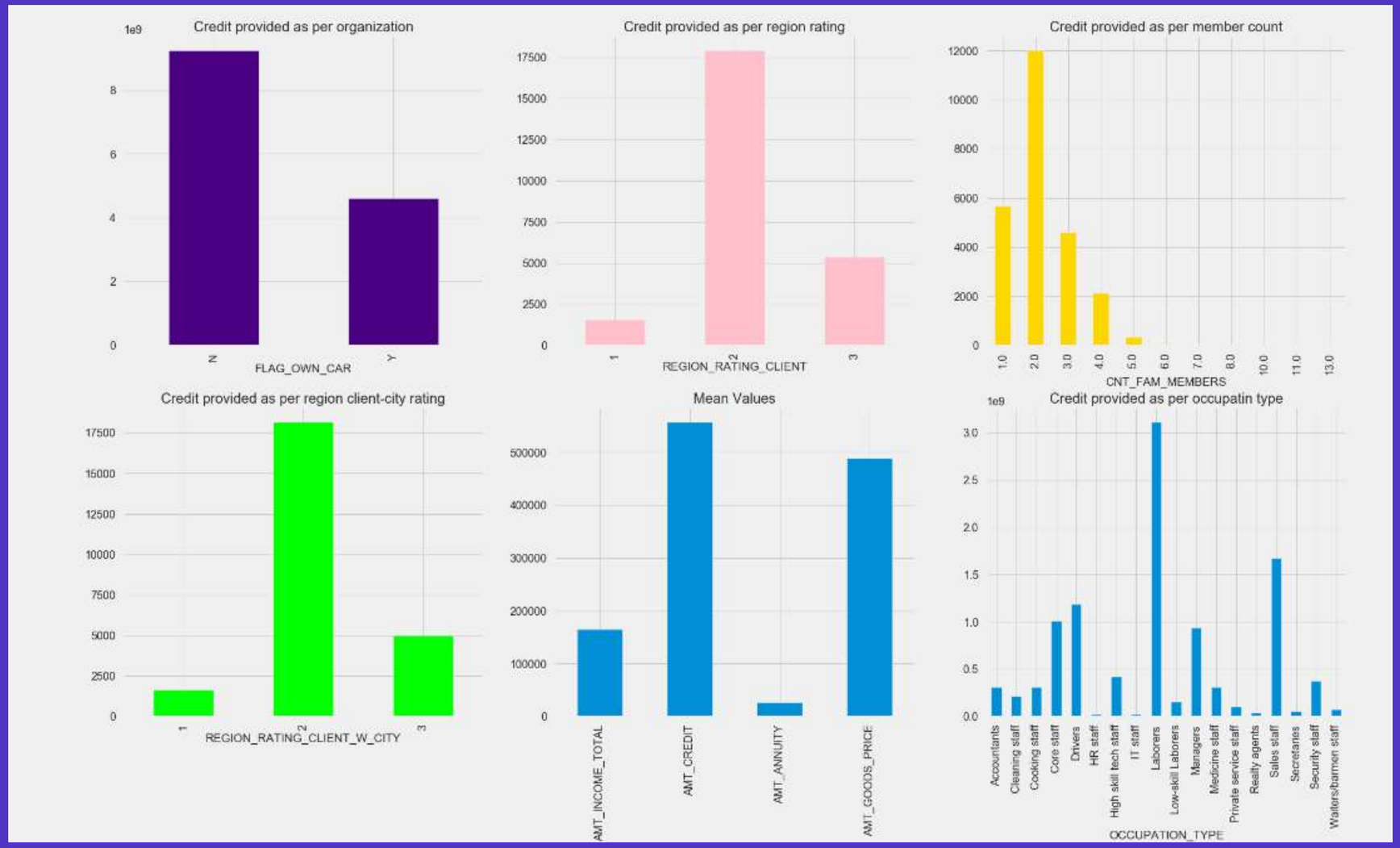
DATA VISUALIZATION

Analyzing Defaulters



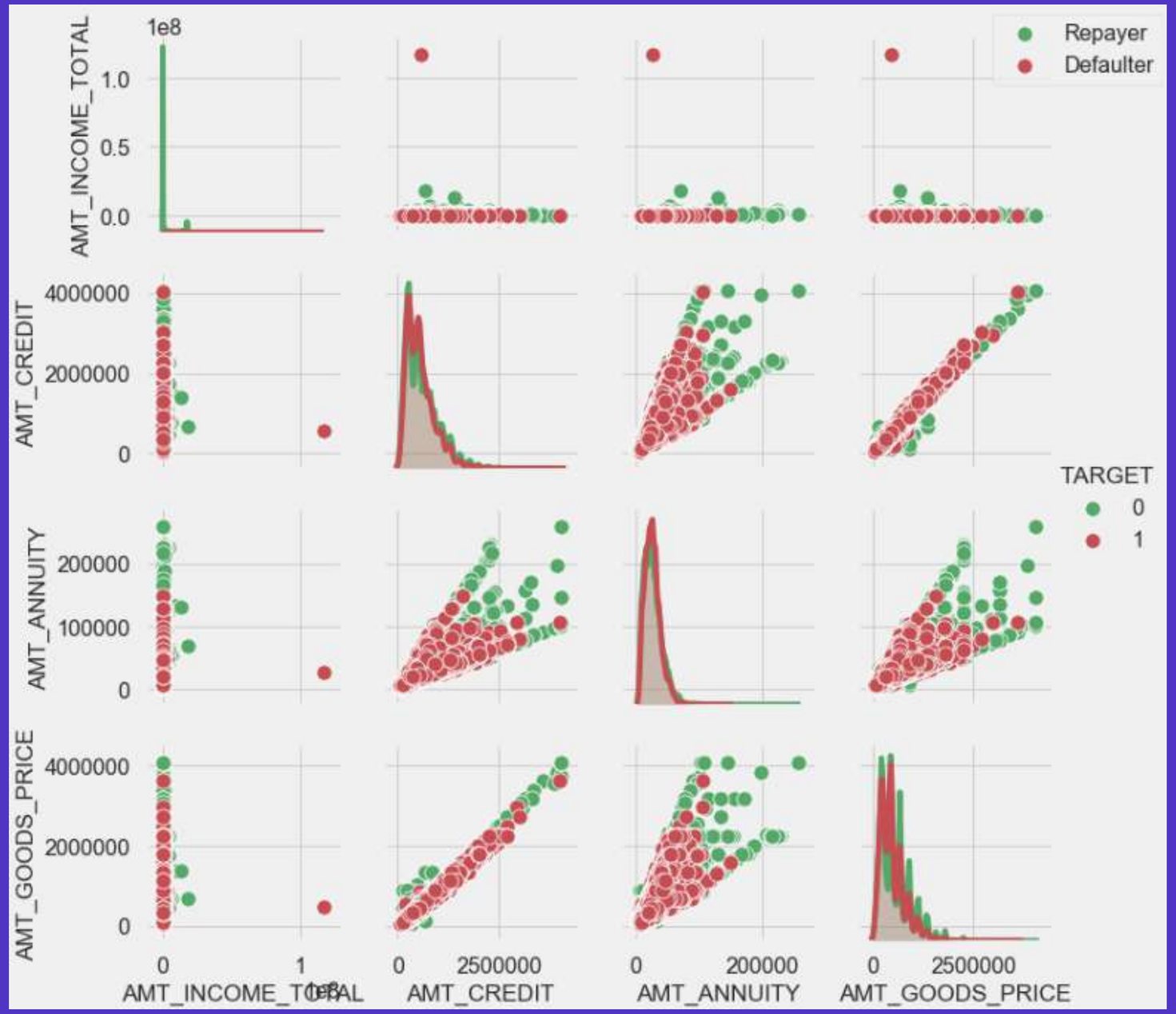
The above visuals are plotted to give us a birds view on the kinds of defaulters. We could find a higher concentration of defaulters in Business Entity type; a greater number of our defaulted clients fall under the female category and are the ones who own a house or an apartment.

DATA VISUALIZATION



Further we could see more defaults have been made in the Cash loans contract type with more than 60% of our defaulted clients being married and about 75% of the defaulted clients have the education type as secondary school. Most of our defaulters do not own a car, have a region rating of 2 and the highest occupation type of our defaulters is "Labourers".

DATA VISUALIZATION



When AMT_ANNUITY >15000 AMT_GOODS_PRICE> 3M, there are a lesser chance of our clients defaulting on the loan repayment. The Credit Amount and the Price of Goods 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 when the Credit amount is greater than 3M.

Data Modelling & Predictive Analytics

DATA MODELLING

By applying the Regression Analysis, we get:

Our Model in Mathematical Form:

AMT_ANNUITY = 26383.838570 + 0.000625*(AMT_INCOME_TOTAL)

For every single unit increase of Total Income there will be 26383.838570 increase in Annuity Amount/ This is a very covariance we have noticed and we ought to reduce this to reduce the number of defaulters.

Model:	OLS	Adj. R-squared:	0.002
Dependent Variable:	AMT_ANNUITY	AIC:	430865.4638
Date:	2022-12-20 15:43	BIC:	430881.2568
No. Observations:	19860	Log-Likelihood:	-2.1543e+05
Df Model:	1	F-statistic:	34.88
Df Residuals:	19858	Prob (F-statistic):	3.56e-09
R-squared:	0.002	Scale:	1.5473e+08

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	26383.8386	90.0126	293.1129	0.0000	26207.4064	26560.2707
AMT_INCOME_TOTAL	0.0006	0.0001	5.9061	0.0000	0.0004	0.0008

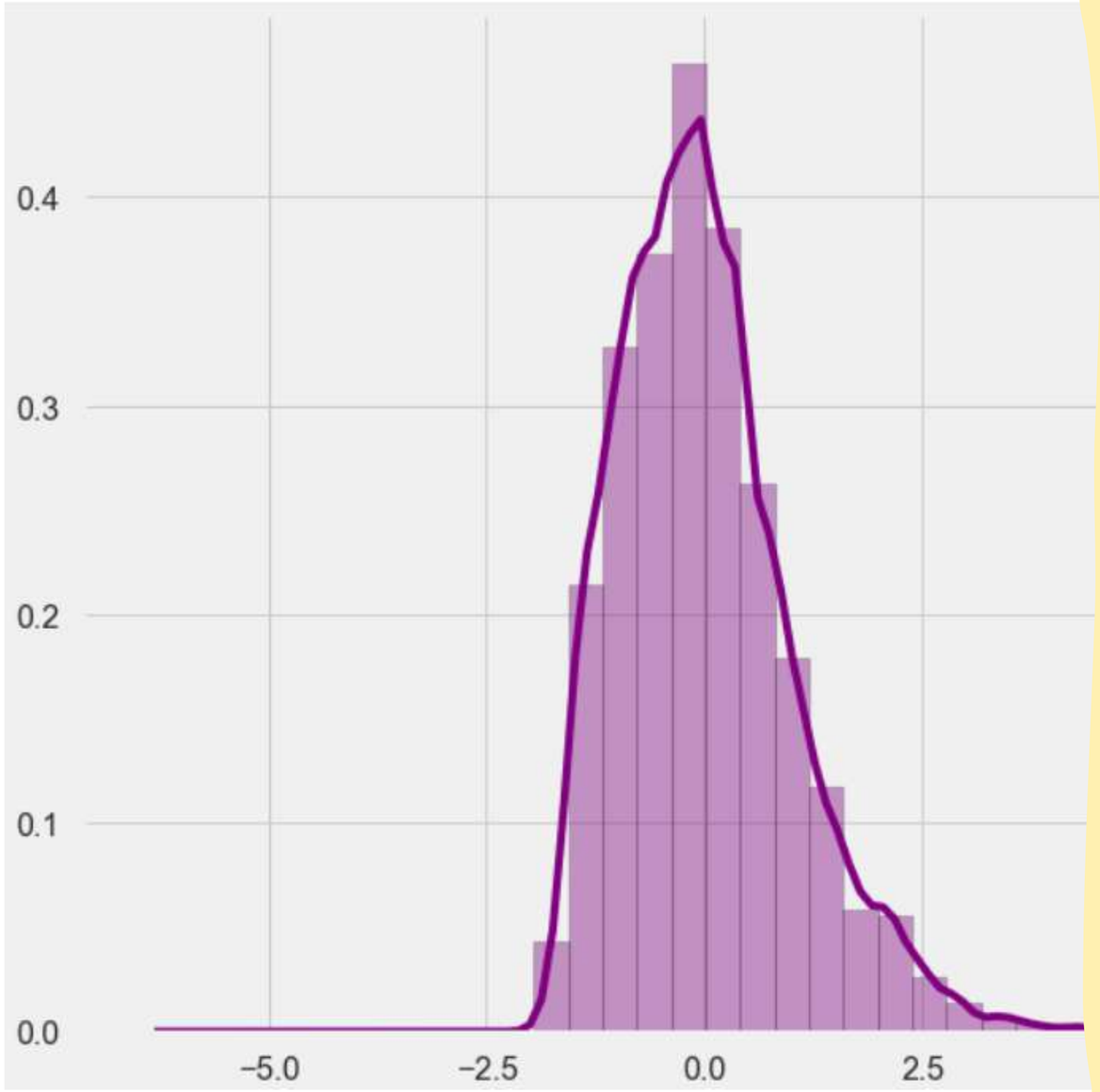
Omnibus:	3477.343	Durbin-Watson:	2.009
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8157.991
Skew:	0.998	Prob(JB):	0.000
Kurtosis:	5.424	Condition No.:	867071

The R-squared is 0.002, which means that the model explains 0.2% of the variance in y. We get this very low variance due to Amount of Income and Annuity being highly subjective like the human nature. Since, β value is not 0 but very close to zero, it means that our model holds statistical significance.

The p-value of t-test of β1 is less than 0.05, so β1 is statistically significant. The p-value of F-test of the model is also less than 0.05, so the model is statistically significant. Hence, we could conclude that our model is statistically significant for analysis.

DATA MODELLING

Our Histogram plot shows a normal distribution.



Summary of our Model

1. Relation from model is :
   AMT_ANNUITY = 26383.838570 + 0.000625* (AMT_INCOME_TOTAL)
2. For every single unit increase of Total Income there will be 26383.838570 increase in Annuity Amount. This is a very covariance we have noticed and we ought to reduce this to reduce the number of defaulters
3. R-Square is 0.002 meaning that 0.2% of variance is explained by the Total Income data on the amount of annuity of the defaulted loan
4. The model is significant, as indicated by p-value(F-Statistic) which is below 0.05
5. The effect of Total Income on Amount of Annuity is significant which is inferred by the p-value of coefficient for Amount of Total Income which is less than 0.05

PREDICTIVE ANALYTICS

The following are the standardized residuals retrieved for the model:

```
Mean of Residuals:  5.265727028633352e-12
Std Dev of Residuals:  12438.546541945676
Mean of Standardized Residuals:  3.4015254527129515e-17
Std Dev of Standardized Residuals:  1.0000000000000013
```

```
R-Square_Train:  0.0018  RMSE_Train:  12438.2334
R-Square_Test:  0.0034  RMSE_Test:  12434.5746

MAE_Train:  97.974
MAE_Test:  97.7233
```

RMSE and MAE Values in Test Data are similar and lesser to that of the RMSE and MAE values in the Train Data. This helps us understand that our regression in not over-fitted. We can also infer that the model's performance in training and Test data sets are similar. The MAE is about 97% which is very high but lesser when compared to other models.

CONCLUSION

One of the primary reasons for loan repayment defaults is because of the Amount of credit being higher to the total income of the client.

Better policies should be undertaken for loan defaulters who are in Maternity Leave, State Servants and unemployed.

Business Entities and Labourers have a high concentration of loan defaults. A high concentration of defaulter count with education type of Secondary School. Hence, we need consider better methodologies on sanctioning loans to the respective client types.

A higher number of loan defaults in Cash loans, hence the bank needs to frame better policies to avoid defaults in cash loan repayment.

Based on our Data Modelling using Regression Analysis, we could statistically conclude that, it would be significant to consider Total Income and Amount of Annuity for developing a predictive analytical model. Although the variance seemed low due to the subjectivity of human nature, we found our model to be statistically significant for analysis.

We see a huge range of interval in the predicted values due to the variation in price of goods, as amount of annuity is calculated on amount of credit which is sanctioned based on price of the goods.

We would advise the bank to sanction loans based on the Total Income of the Clients and revise a few policies to curb and reduce the loan defaults.

Thank you!