EDA Credit Assignment - Contents

1. Problem Statement
2. Assumptions (if any)
3. Approach Overall
4. Relevant Univariate and Bivariate and Segmented Univariate graphs and inferences
5. Conclusions / Results / Recommendations

1.Problem Statement

Suppose you are working for consumer finance company which specializes various types of loans to Urban customers. You are expected to use EDA and domain knowledge to find applicants capable of replaying the loan are not rejected due to the non existence of past credit history and identify applicants who are expected to be defaulters.

This study is aimed at identify patterns that has difficulty paying loan are denied the loan, loan amount decreased, raising the interest rates.

In other words company wants to understand driving factors that will behind loan default or strong variables indicating defaulters.

2.Assumptions (if any)

- ‘XNA’ value is considered as Null values for all variables.
- 40% is set as threshold for null values. ie any column with null values above this threshold will be dropped.
- -ve value for variables considered as absolute, Eg. age of client is taken as positive.

3.Approach Overall

1. Understanding the domain/variables.
2. Import/Load 2 sets of data.
   1. Application data
   2. Previous Application data

Steps for APPLICATION DATA

1. Check the structure/metadata of the data.
2. Missing value check.
3. Outlier Check.
4. Perform Univariate Analysis.
5. Perform Bivariate Analysis.
6. Perform Segmented Univariate. – Segmentation done based on **TARGET** variable.

Steps for PREVIOUS APPLICATION DATA

1. Check the structure/metadata of the data.
2. Missing value check.
3. Outlier Check.
4. Perform Univariate Analysis.
5. Perform Bivariate Analysis.
6. Perform Segmented Univariate. – Segmentation done based on **NAME_CONTRACT_STATUS** variable.

Steps for Merged APPLICATION DATA

1. Merge APPLICATION DATA and PREVIOUS APPLICATION DATA using SK_ID_CURR with inner join.
2. Check the structure/metadata of the data.
3. Missing value check.
4. Outlier Check.
5. Perform Univariate Analysis.
6. Perform Bivariate Analysis.
7. Perform correlation analysis of important variable on TARGET variable.

4.Application Data : Missing value and Outliers Check

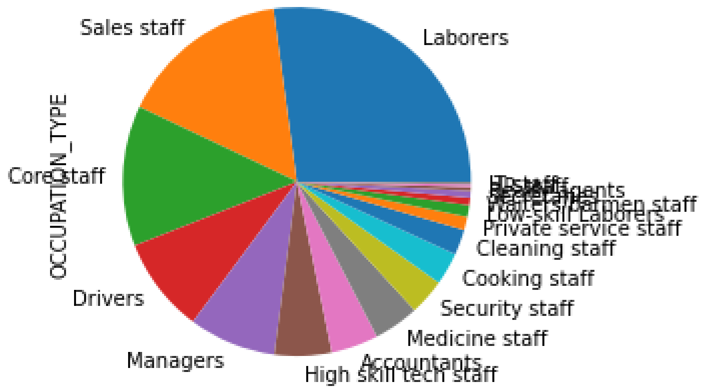
Missing values cleanup - A threshold value of 40% is considered. ie all columns with null values more than 40% are dropped. After this cleanup total 73 columns remained in the data set.

Outliers are identified for many numerical columns AMT_INCOME_TOTAL, AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE using Box Plots.

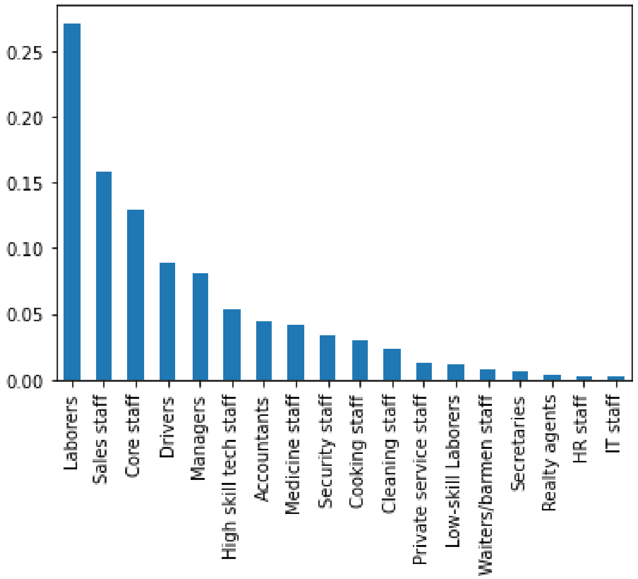
CNT_CHILDREN has many values, we can categorize into fixed categories, NONE for 0, ONE for 1, TWO for 2, OTHERS for rest.
we will add a column PARENT_CAT with these values and drop CNT_CHILDREN.

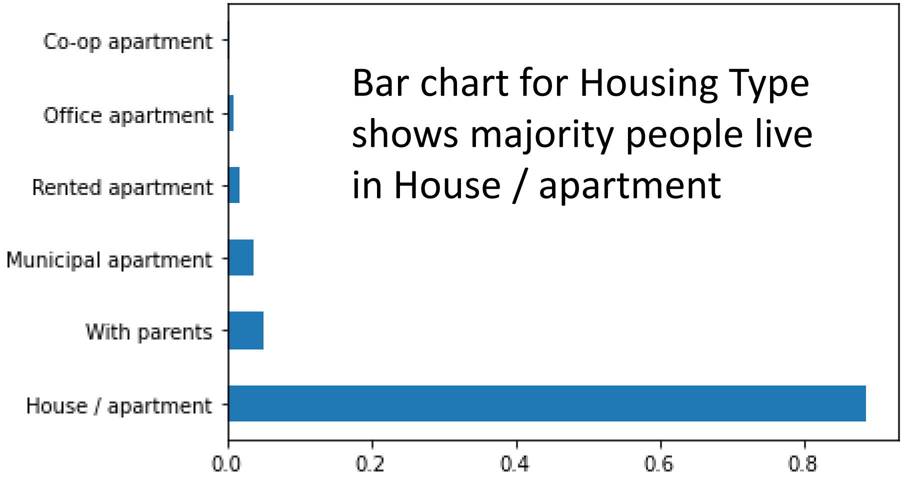
OCCUPATION_TYPE has many null values, null values are replaced by category 'OTHERS' so that it doesn't miss out during analysis.

4.Application Data : Univariate Analysis

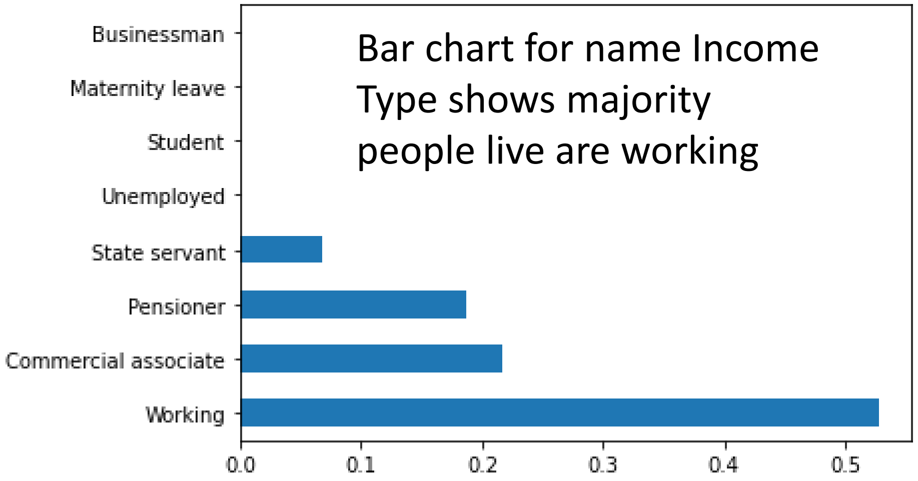


Pie Chart of Occupation Type shows that Most of applications are of Occupation Type Laborers followed by Sales Staff, Core staff, Drivers etc



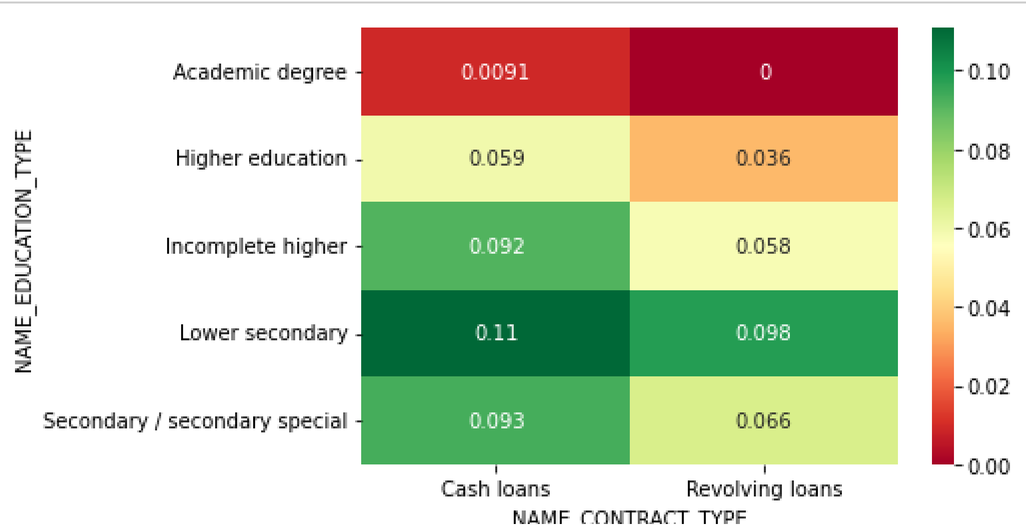


Bar chart for Housing Type shows majority people live in House / apartment



Bar chart for name Income Type shows majority people live are working

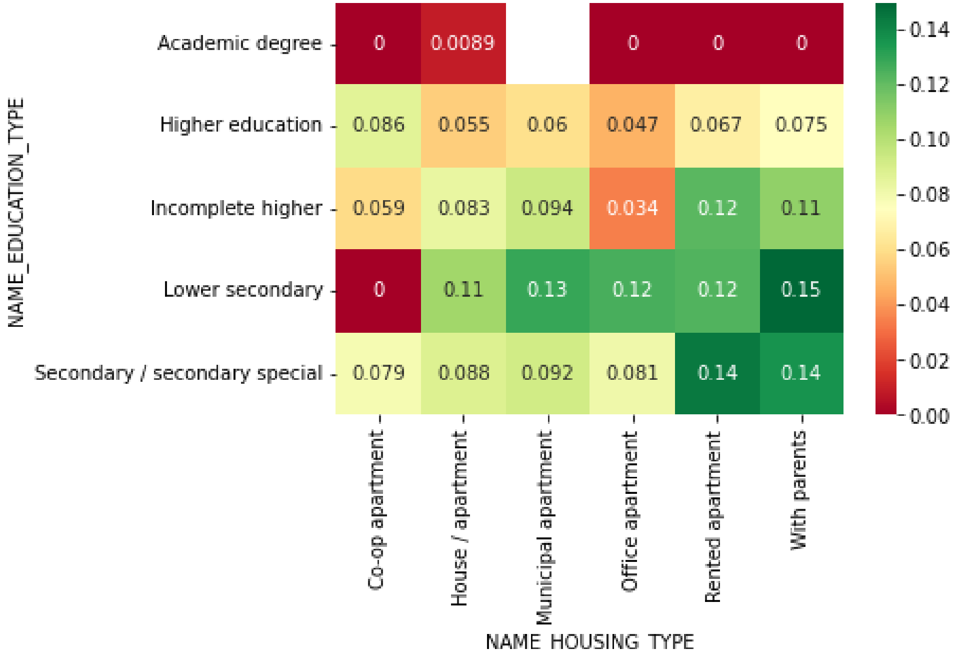
4.Application Data : Bivariate Analysis



create heat map of NAME_EDUCATION_TYPE vs NAME_CONTRACT_TYPE vs TARGET

We see below patterns for Defaulters ie Target value = 1

- lower secondary with Cash Loans
- Lower secondary with revolving loans
- Secondary/secondary special with Cash loans.

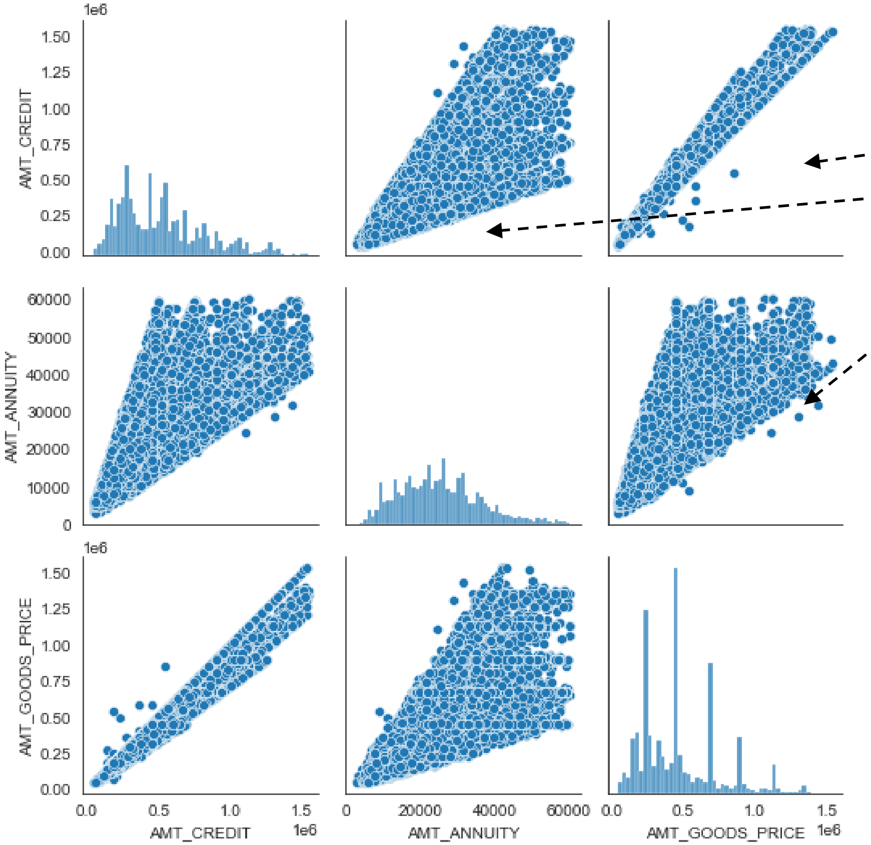


create heat map of NAME_EDUCATION_TYPE vs NAME_HOUSING_TYPE vs TARGET

We see below patterns for Defaulters ie Target value = 1

- Lower secondary with Parents
- Secondary/secondary special with Rented Apartment
- Lower secondary with Municipal Apartment

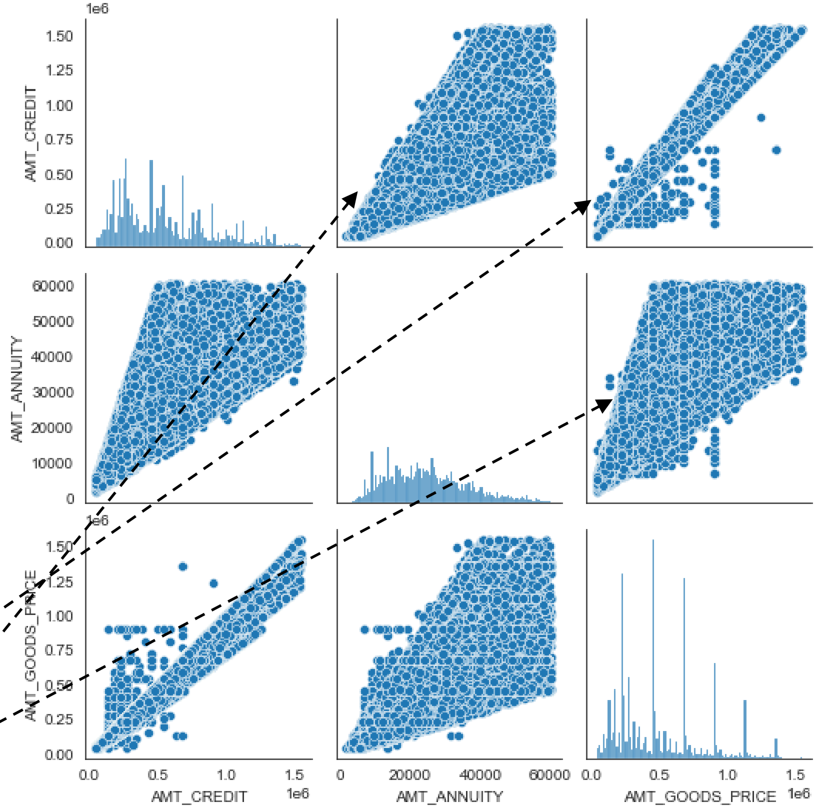
4.Application Data : Segmented Univariate Analysis



Create pair plot for 'AMT_CREDIT', 'AMT_ANNUITY','AMT_GOODS_PRICE' for Defaulters ie Target = 1 , Defaulters

It observed clear pattern as below

- 'AMT_CREDIT' and 'AMT_GOODS_PRICE'
- 'AMT_CREDIT' and 'AMT_ANNUITY'
- 'AMT_ANNUITY' and 'AMT_GOODS_PRICE'

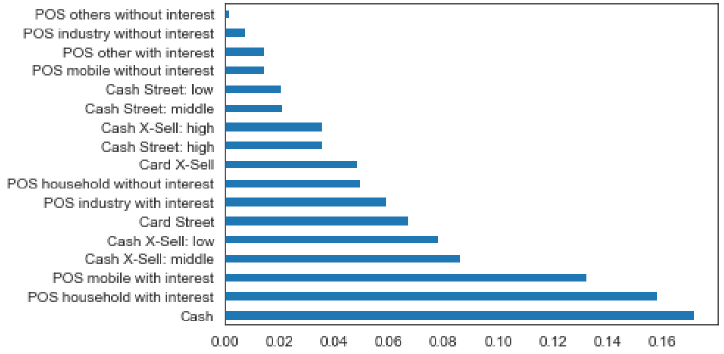


Create pair plot for 'AMT_CREDIT', 'AMT_ANNUITY','AMT_GOODS_PRICE' for Defaulters ie Target = 0 , Non Defaulters

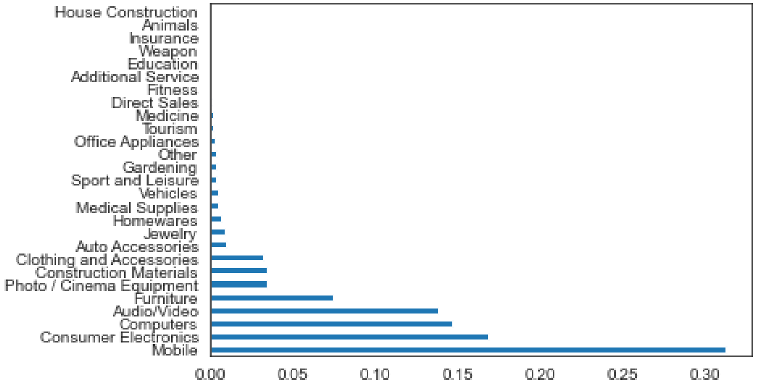
It observed clear pattern as below

- 'AMT_CREDIT' and 'AMT_GOODS_PRICE'
- 'AMT_CREDIT' and 'AMT_ANNUITY'
- 'AMT_ANNUITY' and 'AMT_GOODS_PRICE'

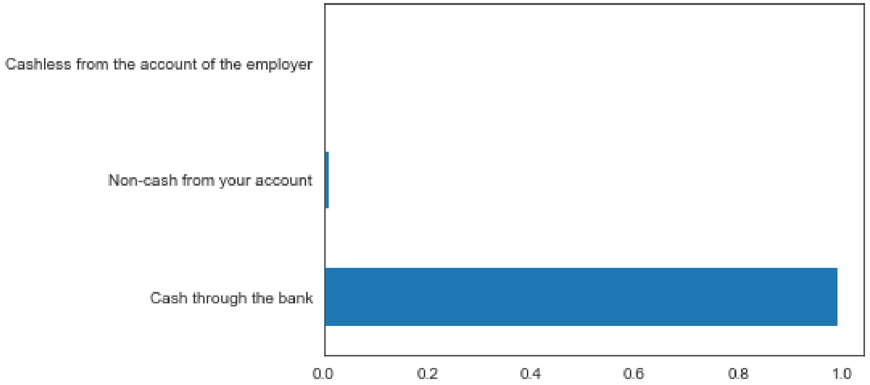
4.Previous Application Data : Univariate Analysis



PRODUCT_COMBINATION is found mostly Cash, followed by POST household with interest

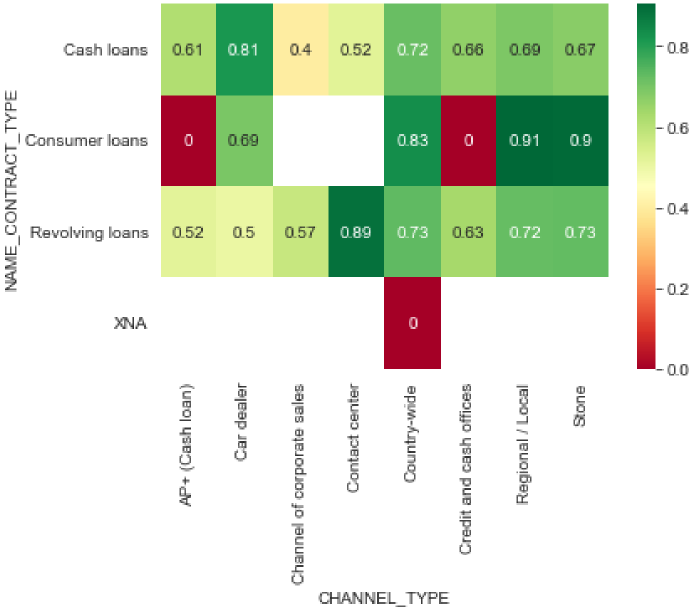


NAME_GOODS_CATEGORY is 35% applied for mobile followed 17% for consumer electronics and 15% for computers.



NAME_GOODS_CATEGORY is 35% applied for mobile followed 17% for consumer electronics and 15% for computers.

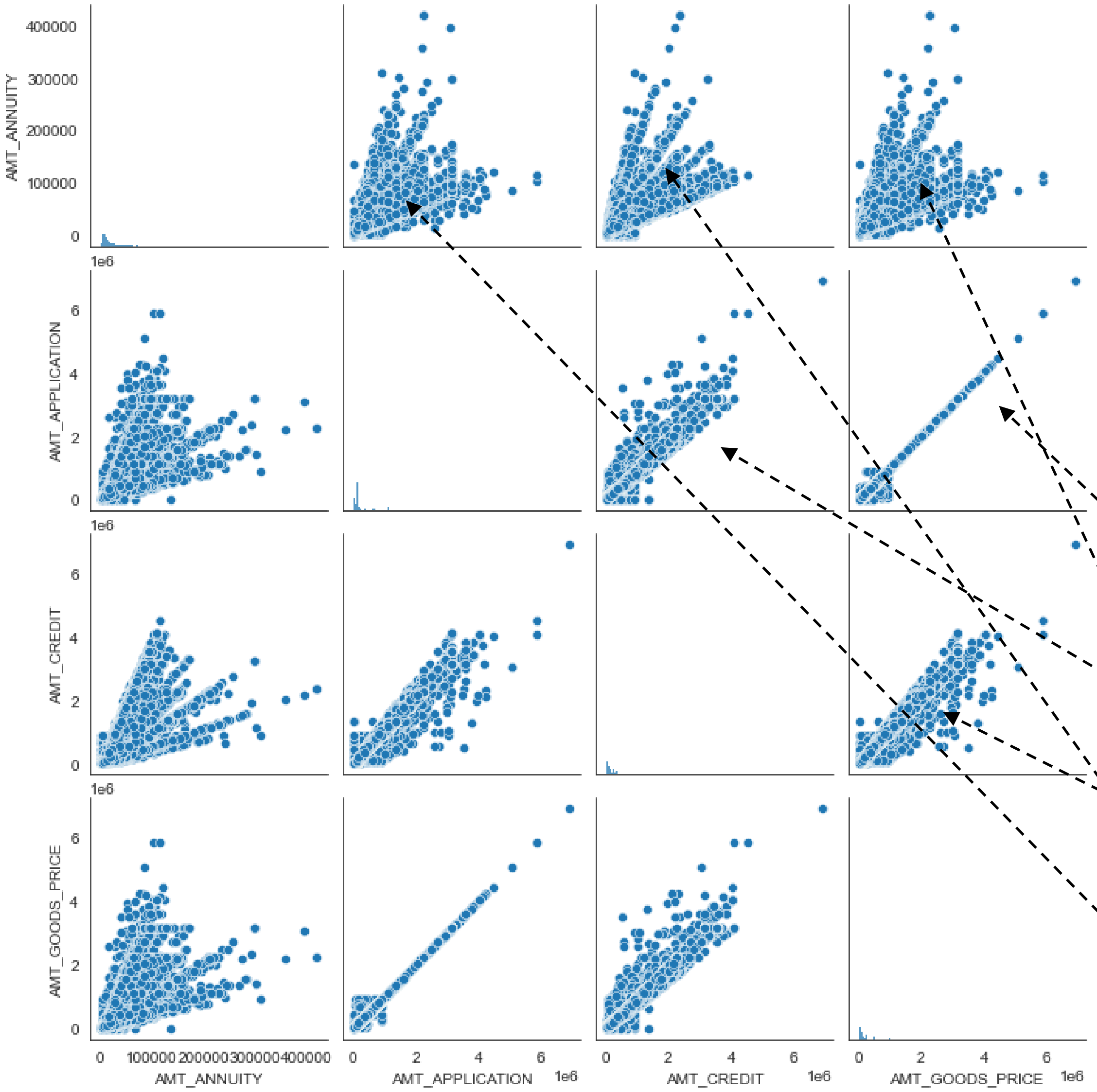
4.Previous Application Data : Bivariate Analysis



create heat map of NAME_CONTRACT_TYPE vs CHANNEL_TYPE to see correlation with TARGET

It is seen there is correlation b/n contract types Consumer loans and Regional / Local and channel types consumer loans and Stone.

4.Previous Application Data : Bivariate Analysis



Create pair plot to understand correlation b/n 'NAME_CONTRACT_STATUS', 'AMT_ANNUITY', 'AMT_APPLICATION', 'AMT_CREDIT', 'AMT_GOODS_PRICE'

By looking at pair plots we can make conclusion,
Clear correction can be seen for below

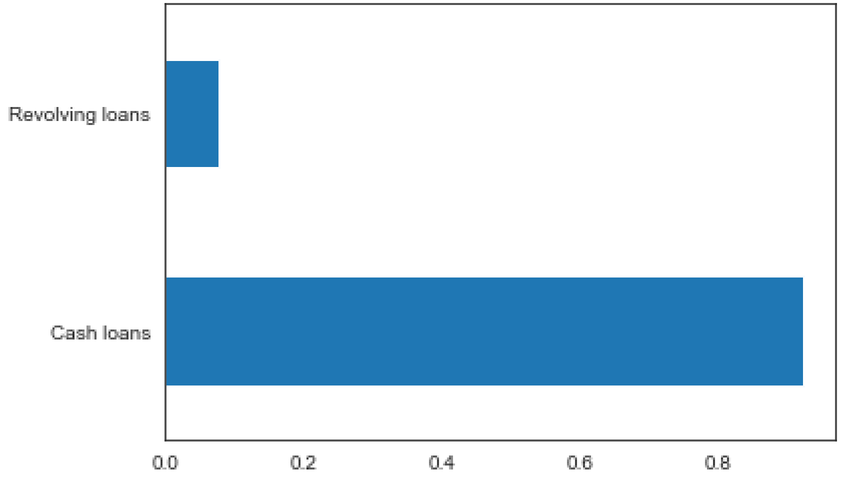
-AMT_GOODS_PRICE vs AMT_APPLICATION,

-AMT_CREDIT vs AMT_APPLICATION,

-AMT_GOODS_PRICE vs AMT_CREDIT

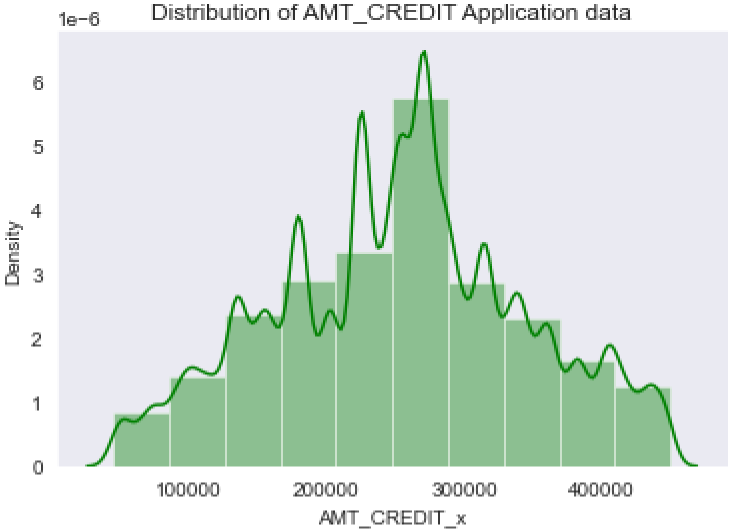
Also correlation for below
-'AMT_ANNUITY' vs 'AMT_GOODS_PRICE'
- 'AMT_CREDIT', vs 'AMT_ANNUITY'
- 'AMT_ANNUITY' vs 'AMT_APPLICATION'

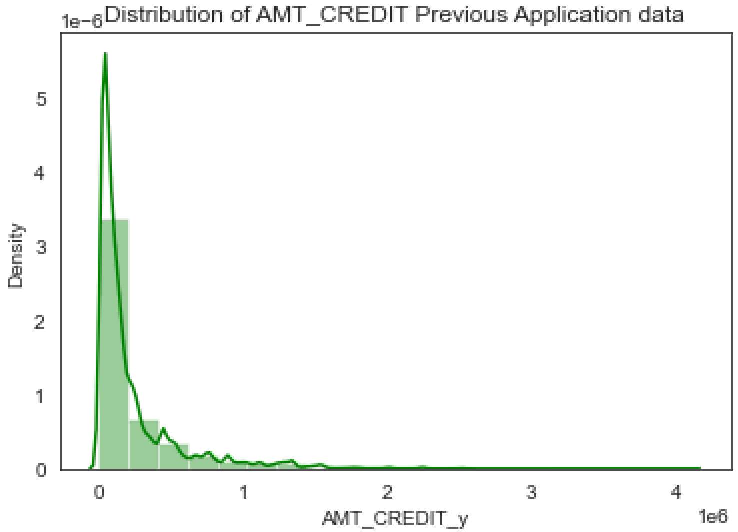
Merged Data Set : Univariate Analysis



NAME_CONTRACT_TYPE_x

- ➢ 85% - Cash Loans
- ➢ 15% - Revolving Loans

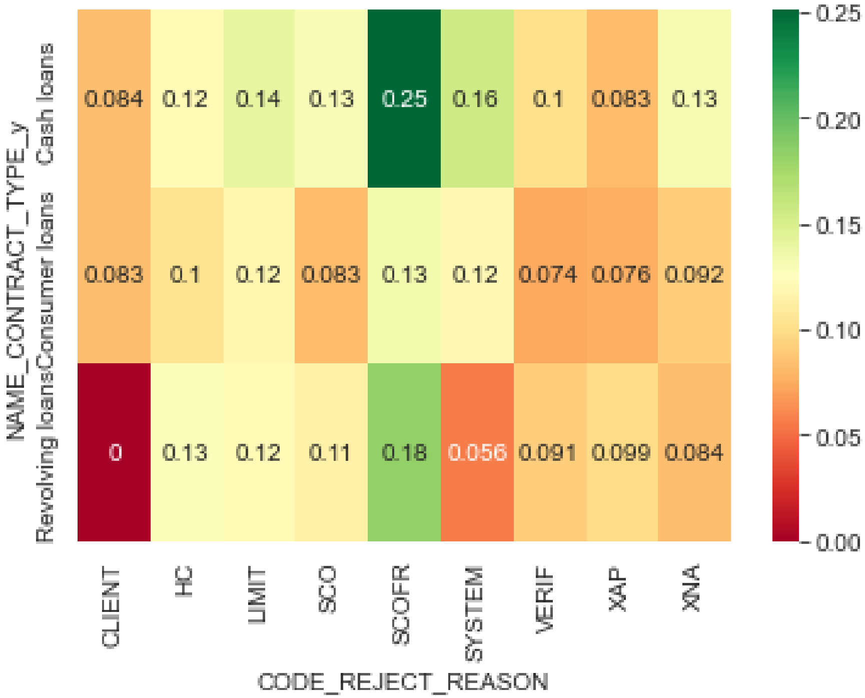




Comparison of AMT_CREDIT distribution for Application Data vs Previous Application data

- It is observed Mean of AMT_CREDIT for Application data is 5.8 Lacs, whereas Mean of AMT_CREDIT for Previous application data is around 3 Lavs

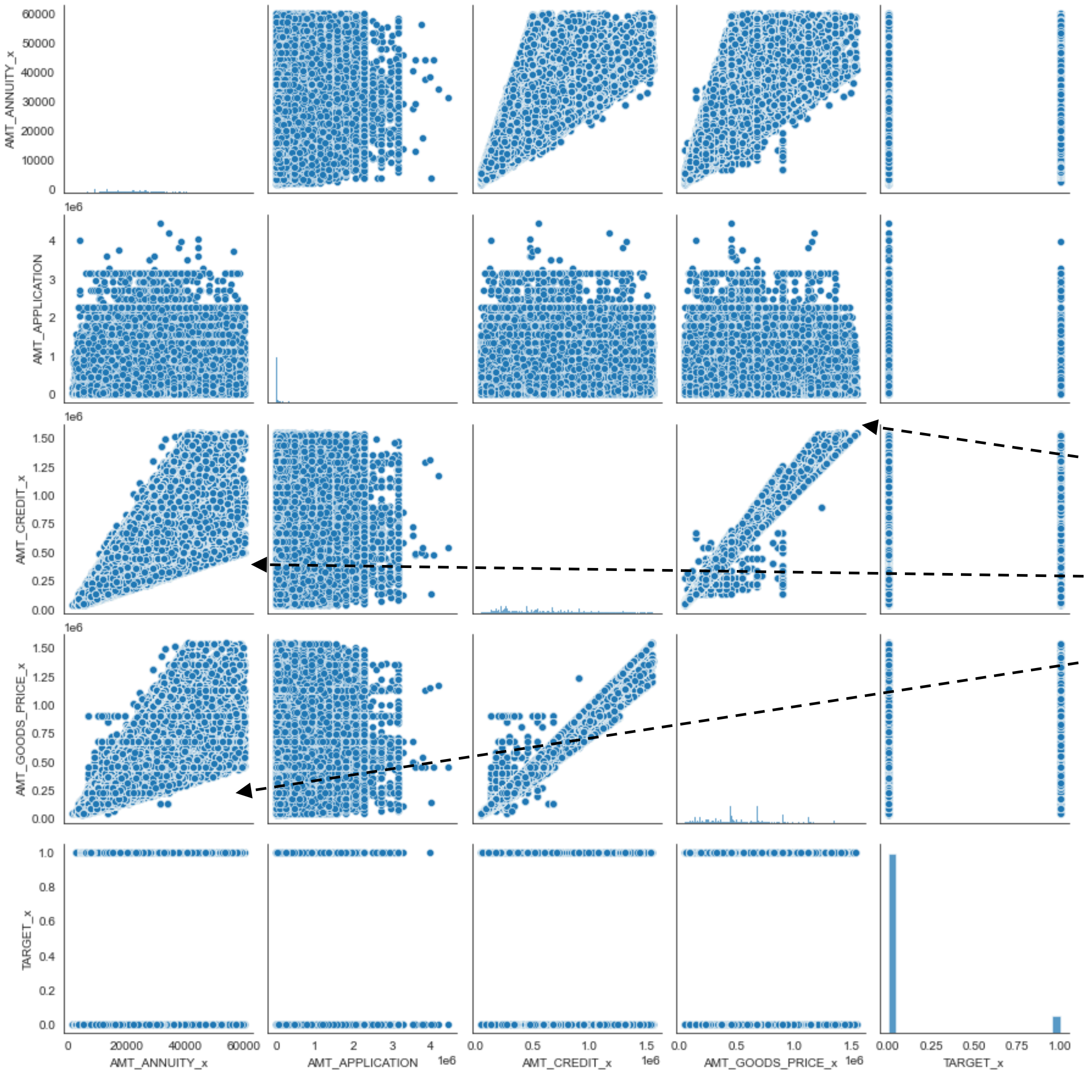
Merged Data Set : Bivariate Analysis -



create heat map of NAME_CONTRACT_TYPE_y, CODE_REJECT_REASON with TARGET to see correlation with TARGET

from heatmap, it is clear that there is correlation b/n CONTRACT_TYPE Cash Loans and Reject reason SCOFR to defaulters, followed by Cash loans vs SCOFR

Merged Data Set : Pairplot using variables



Pairplot made using 'NAME_CONTRACT_STATUS', 'AMT_ANNUITY_x', 'AMT_APPLICATION', 'AMT_CREDIT_x', 'AMT_GOODS_PRICE_x', 'TARGET_x'

Below observations found from this pairplot.

- Clear positive correlation b/n AMT_CREDIT_x and AMT_GOODS_PRICE_x,
- Positive correlation b/n AMT_ANNUITY_x and AMT_CREDIT_x
- Positive correlation b/n AMT_GOODS_PRICE_x and AMT_ANNUITY_x

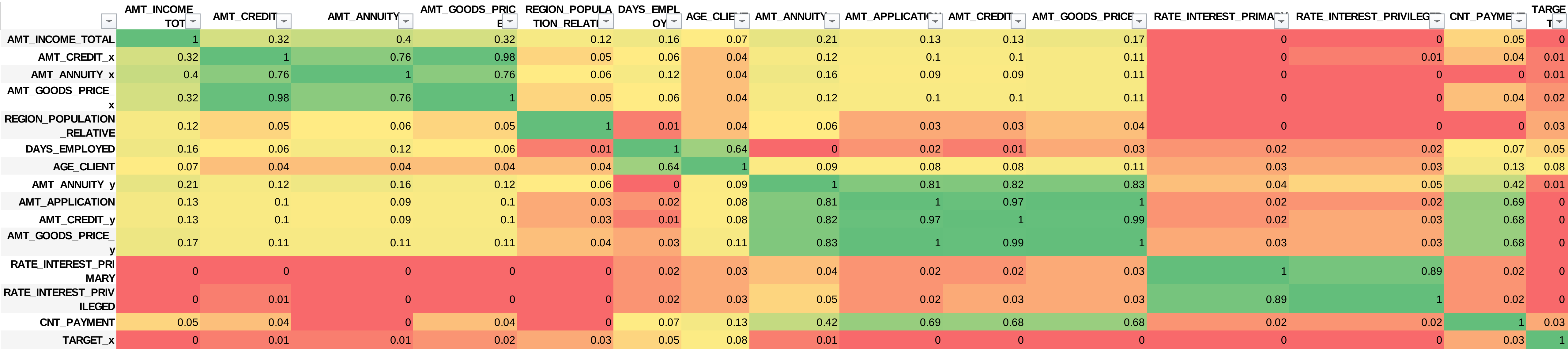
Merged Data Set : Correlation chart of important variable with TARGET

Correlation chart made using below Important Factors

- AMT_INCOME_TOTAL
- AMT_CREDIT_x
- AMT_ANNUITY_x
- AMT_GOODS_PRICE_x
- REGION_POPULATION_RELATIVE
- DAYS_EMPLOYED
- AGE_CLIENT
- AMT_ANNUITY_y
- AMT_APPLICATION
- AMT_CREDIT_y
- AMT_GOODS_PRICE_y
- RATE_INTEREST_PRIMARY
- RATE_INTEREST_PRIVILEGED
- CNT_PAYMENT

8 Important Factors impacting TARGET value

1.00	AGE_CLIENT
2.00	DAYS_EMPLOYED
3.00	REGION_POPULATION_RELATIVE
3.00	CNT_PAYMENT
4.00	AMT_GOODS_PRICE_x
5.00	AMT_CREDIT_x
5.00	AMT_ANNUITY_x
5.00	AMT_ANNUITY_y



	AMT_INCOME TOT	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRIC E	REGION_POPULA TION_RELATI	DAYS_EMPL OY	AGE_CLIE	AMT_ANNUITY	AMT_APPLICATI	AMT_CREDIT	AMT_GOODS_PRICE	RATE_INTEREST_PRIMA	RATE_INTEREST_PRIVILEG	CNT_PAYME	TARGE T
AMT_INCOME_TOTAL	1	0.32	0.4	0.32	0.12	0.16	0.07	0.21	0.13	0.13	0.17	0	0	0.05	0
AMT_CREDIT_x	0.32	1	0.76	0.98	0.05	0.06	0.04	0.12	0.1	0.1	0.11	0	0.01	0.04	0.01
AMT_ANNUITY_x	0.4	0.76	1	0.76	0.06	0.12	0.04	0.16	0.09	0.09	0.11	0	0	0	0.01
AMT_GOODS_PRICE_x	0.32	0.98	0.76	1	0.05	0.06	0.04	0.12	0.1	0.1	0.11	0	0	0.04	0.02
REGION_POPULATION_RELATIVE	0.12	0.05	0.06	0.05	1	0.01	0.04	0.06	0.03	0.03	0.04	0	0	0	0.03
DAYS_EMPLOYED	0.16	0.06	0.12	0.06	0.01	1	0.64	0	0.02	0.01	0.03	0.02	0.02	0.07	0.05
AGE_CLIENT	0.07	0.04	0.04	0.04	0.04	0.64	1	0.09	0.08	0.08	0.11	0.03	0.03	0.13	0.08
AMT_ANNUITY_y	0.21	0.12	0.16	0.12	0.06	0	0.09	1	0.81	0.82	0.83	0.04	0.05	0.42	0.01
AMT_APPLICATION	0.13	0.1	0.09	0.1	0.03	0.02	0.08	0.81	1	0.97	1	0.02	0.02	0.69	0
AMT_CREDIT_y	0.13	0.1	0.09	0.1	0.03	0.01	0.08	0.82	0.97	1	0.99	0.02	0.03	0.68	0
AMT_GOODS_PRICE_y	0.17	0.11	0.11	0.11	0.04	0.03	0.11	0.83	1	0.99	1	0.03	0.03	0.68	0
RATE_INTEREST_PRIMARY	0	0	0	0	0	0.02	0.03	0.04	0.02	0.02	0.03	1	0.89	0.02	0
RATE_INTEREST_PRIVILEGED	0	0.01	0	0	0	0.02	0.03	0.05	0.02	0.03	0.03	0.89	1	0.02	0
CNT_PAYMENT	0.05	0.04	0	0.04	0	0.07	0.13	0.42	0.69	0.68	0.68	0.02	0.02	1	0.03
TARGET_x	0	0.01	0.01	0.02	0.03	0.05	0.08	0.01	0	0	0	0	0	0.03	1

5.Conclusions / Results / Recommendations - 1

One key conclusion after the analysis is that 8 Important Factors that influences defaulters ie TARGET = 1 in the below order.

1.00	AGE_CLIENT
2.00	DAYS_EMPLOYED
3.00	REGION_POPULATION_RELATIVE
3.00	CNT_PAYMENT
4.00	AMT_GOODS_PRICE_x
5.00	AMT_CREDIT_x
5.00	AMT_ANNUITY_x
5.00	AMT_ANNUITY_y

We can evaluate these factors for new applicants and come up with weighted score and decide if he is eligible for the loan or not. If he does not meet all the criteria meet but very close, we can increase the interest rate or we reduce the loan amount for him or increase down payment.

Using pairplot, heatmap we can identify the pattern which variables are closely correlated.

Also we can make many decisions on which category of customers should be given priority based on univariate analysis of categorical variable and numerical variables.

Also we can use this analysis in portfolio and risk analysis of financial products.