Credit EDA Case Study Analysis

VARUN PATEL
HARSHA BELURKAR

Exploratory Data Analysis : Overview

OBJECTIVE → DATA UNDERSTANDING → DATA ANALYSIS → INFERENCES

OBJECTIVE

Loan approval based on the applicant's profile.
1. Likely to repay the loan, then not approving the loan results in a loss of business to the company
2. If the applicant is not likely to repay the loan, may lead to financial loss of company

DATA UNDERSTANDING

Steps Involved :
- Import the Data
- Data Imputation
- Outlier Analysis
- Inspecting datatypes

DATA ANALYSIS

Steps Involved:
- Data Imbalance
- Univariate Analysis
- Bivariate Analysis
- Correlation

INFERENCES

Obtaining Insights based on Results from Data Understanding and Data Analysis and Providing recommendations for the company to approve or reject the loans

Data Understanding

Dataset used : Application.csv

We have imported necessary libraries required
Inspected dataset using Shape, Info, Describe and head functions

Shape of the data frame : (307511, 122)
Total no of columns with missing values > 50% : 41

Dropped the Columns which are populated less than 50 %, as these might not be helpful in giving the right understanding of the data.
Retained only the important columns that are found to be relevant for the analysis.

Import all the necessary libraries

```
#Importing Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
# Supress Warnings
import warnings
warnings.filterwarnings('ignore')
```

```
# Create a class color for setting print formatting
class color:
   BLUE = '\033[94m'
   BOLD = '\033[1m'
   END = '\033[0m'
```

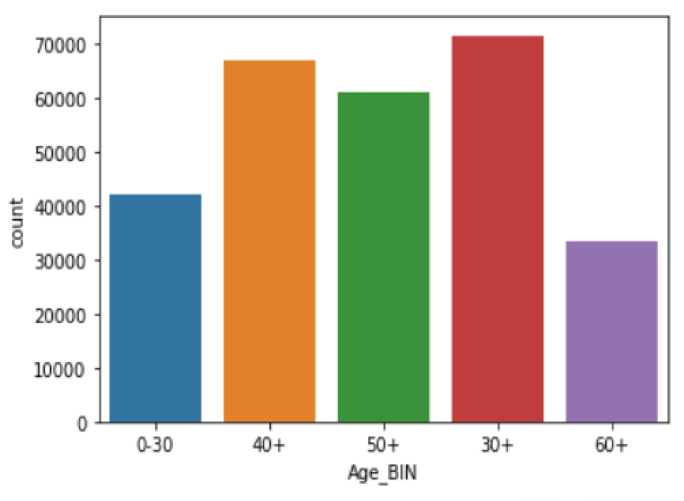
```
df.info()
```

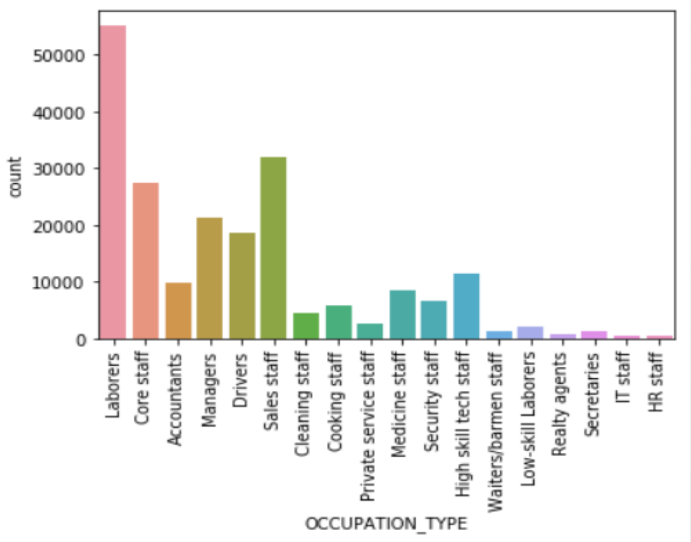
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Columns: 122 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB
```

Data Imputation

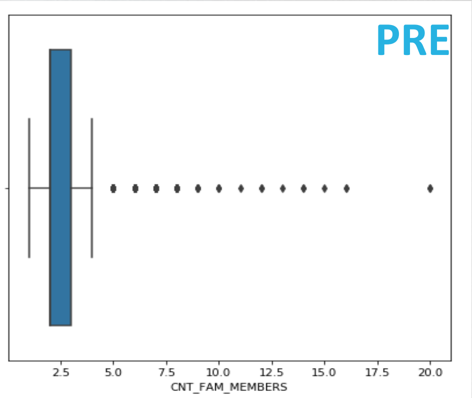
Data imputation is used for replacing missing values via statistical computation, binning them with appropriate values

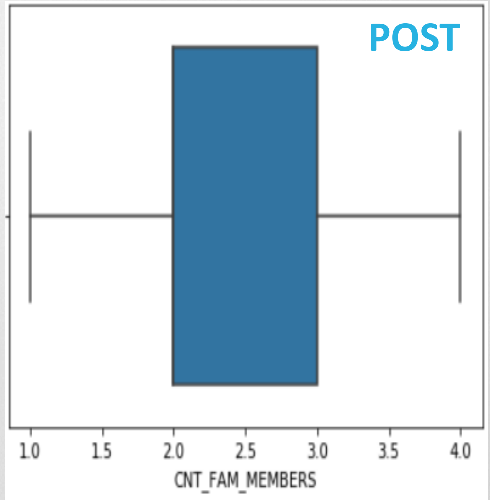
- Categorical Variables: Used MODE for the imputation
  E.g.: OCCUPATION_TYPE , ORGANIZATION_TYPE and CODE_GENDER
- Quantitative Variables: Used Median / Mode for the imputation based on outliers.
  E.g.: AMT_ANNUITY, AMS_GOOD_PRICE and CNT_FAM_MEMBERS
- Binned the Continuous Variables
- Outliers Identification and Treatment – Chose the columns using Box plot
  E.g.: CNT_CHILDREN, AMT_INCOME_TOTAL and CNT_FAM_MEMBERS.





Outliers Treatment – (CNT_FAM_MEMBERS)





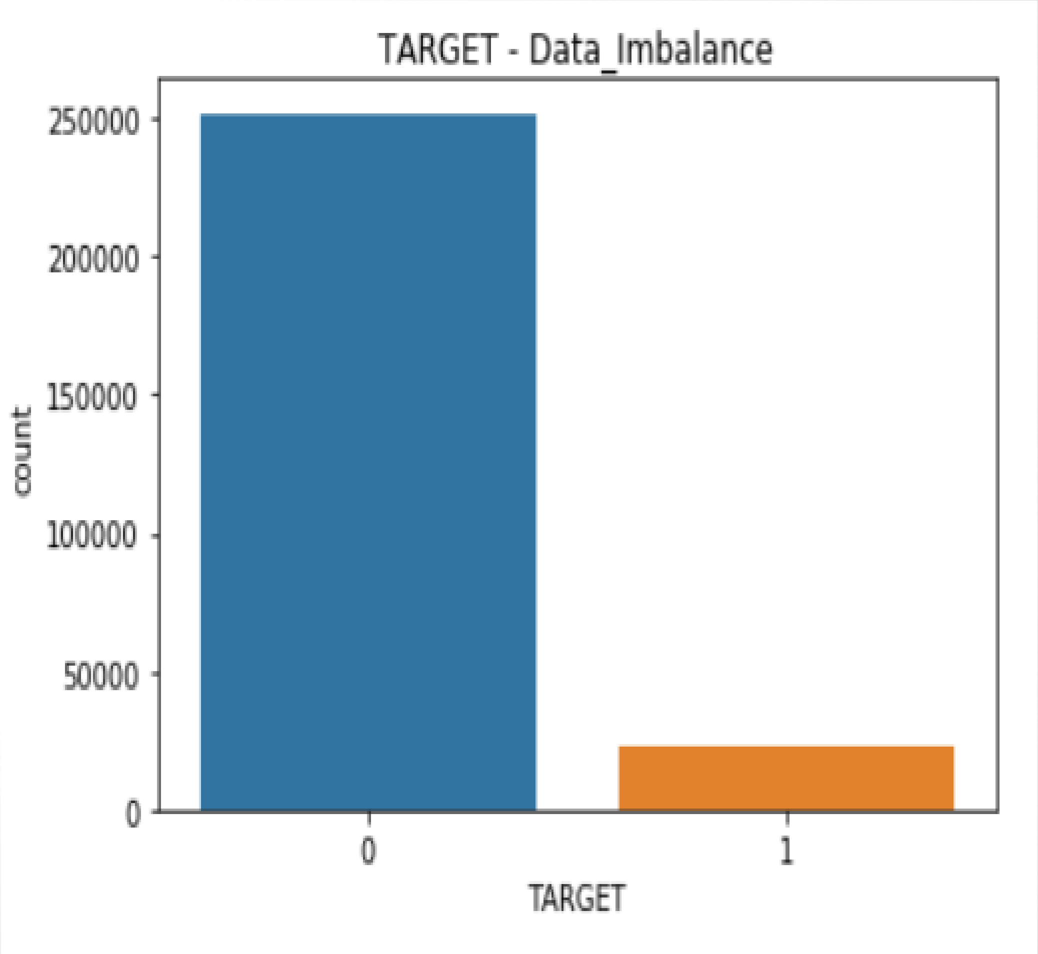
Data Analysis : Data Imbalance

Data Imbalance checks w.r.t TARGET:

Divided the data frame into Target 0 having payment difficult difficulty and Target 1 as all other types

- Clients with difficulty - Target 1 : 22919
- Clients with No difficulty - Target 0 : 251686
- Percentage data imbalance with difficulty(%) : 8.35
- Percentage data imbalance with no difficulty(%): 91.65
- Ratio of imbalance: 10.982 : 1

Data Imbalance in the application data is found and its hugely leaned towards the customers with out having any difficulty to repay the loan i.e., 91.65% of the applicants doesn't find any difficulty in the repayment of the loan

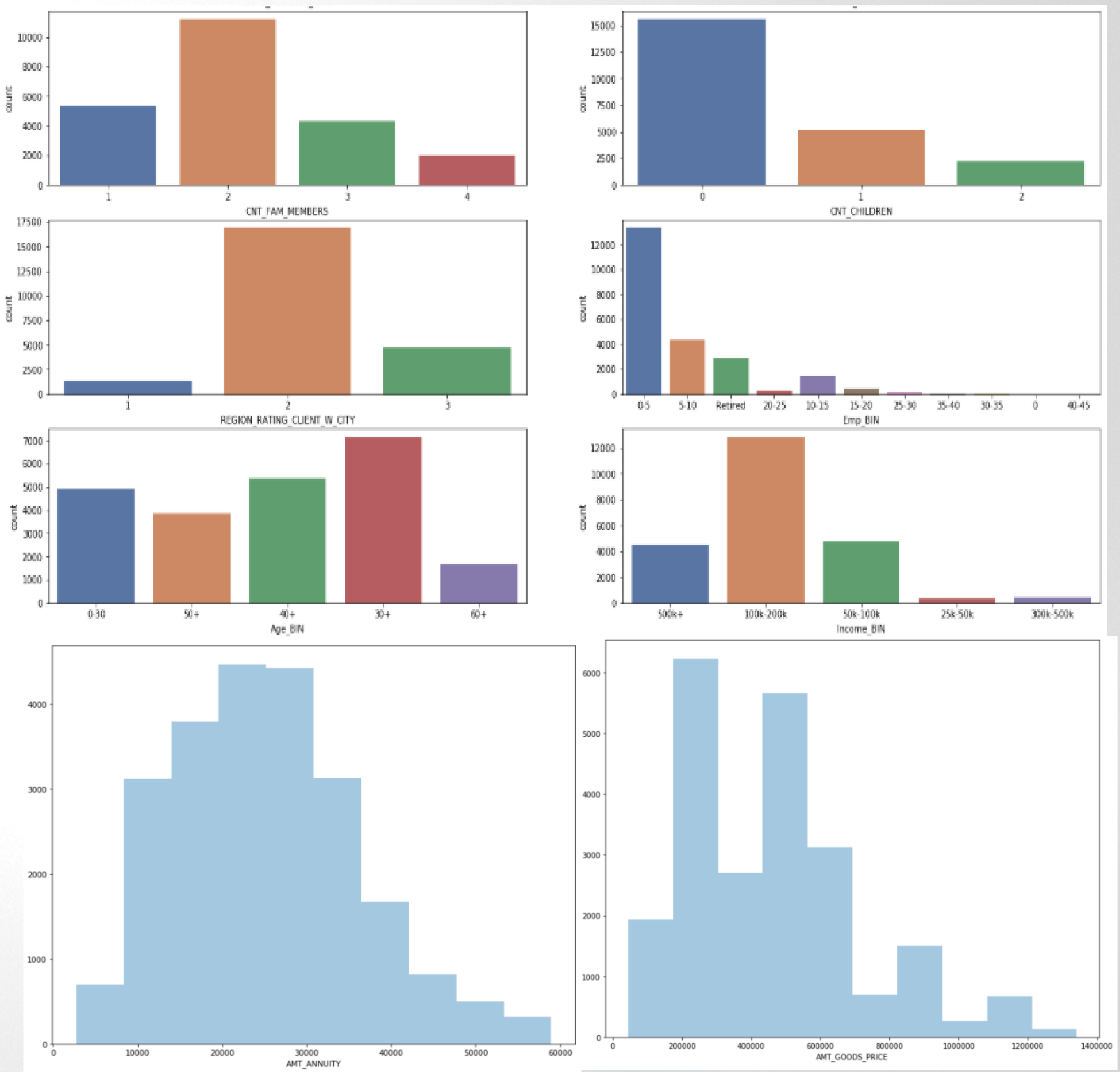


Data Analysis: Univariate Analysis

<u>Observations for Target 0 and Target 1 Univariate Analysis:</u>

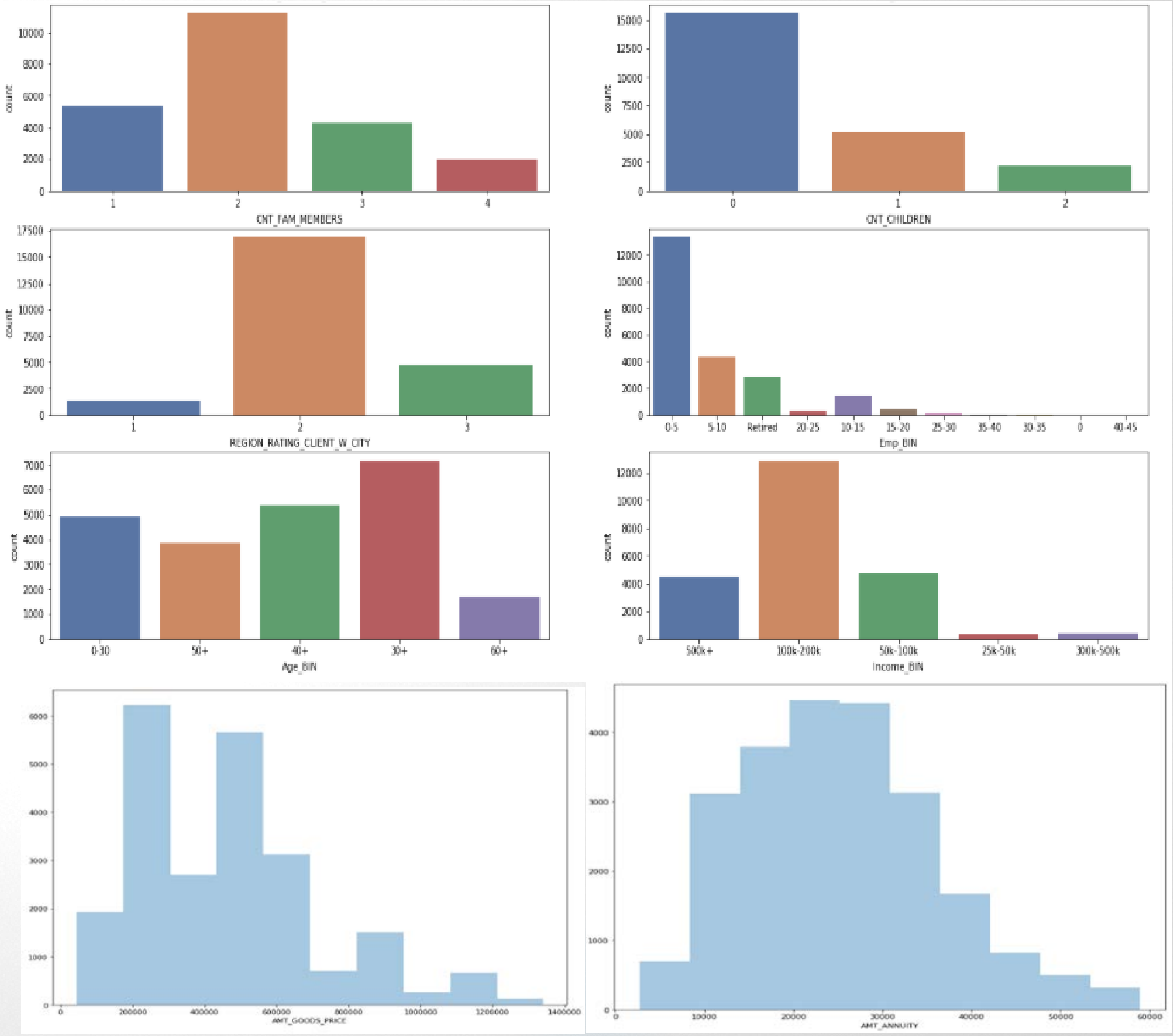
- Loan defaults proportion is less for Married people compare to non-defaulters
- But It seems for Single/Civil Married customers, the loan defaulter proportion is little higher.
- Its seems customer living with Parents have little more proportion of defaulting compared to non-defaulters
- Similarly Municipal and Rented apartment accommodation shows slightly higher proportion towards defaulting
- Its seems customers who are currently working have higher proportion of defaulters
- Pensioners seems to be pay back loan , so their proportion is less on defaulters
- Its seems customers with profession as Laborer have higher proportion of defaulters

Data Set: Applicant who don't find it difficult to repay loan TARGET:0



- Another observation is as IT/HR staff have lower proportion of defaulting
- Customers with Secondary education have higher proportion of defaulting if compared to non-defaulters
- The income of the customers seems to have similar distribution for both defaulters and non-defaulters
- The Average income seems to be around 140K for both segments
- The defaulters seems to have more outliers compared to non-defaulters
- The average annuity is similar for both defaulters and non defaulters around 30K
- The median age for defaulters are around 14000 days older which would be around 40 Years
- It looks like as the age increases proportion of defaulters decreases
- The younger customers seems to have higher proportion of defaulters

Data Set: Applicant who find it difficult to repay loan TARGET:1



Data Analysis: Bivariate Analysis

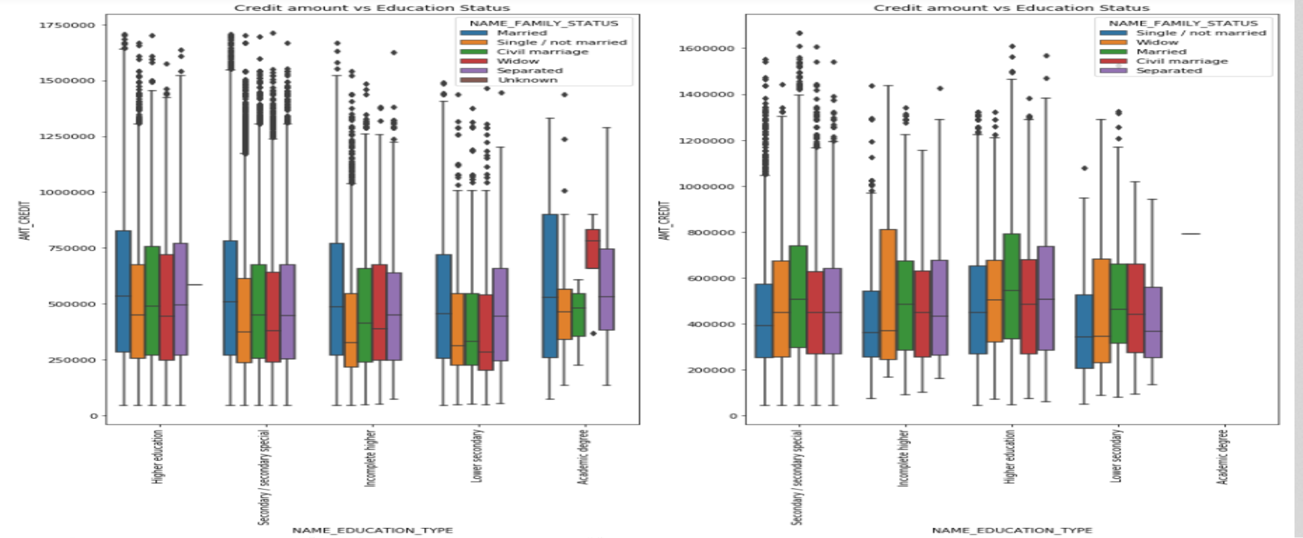
For below variables we have computed correlation between 2 variables :

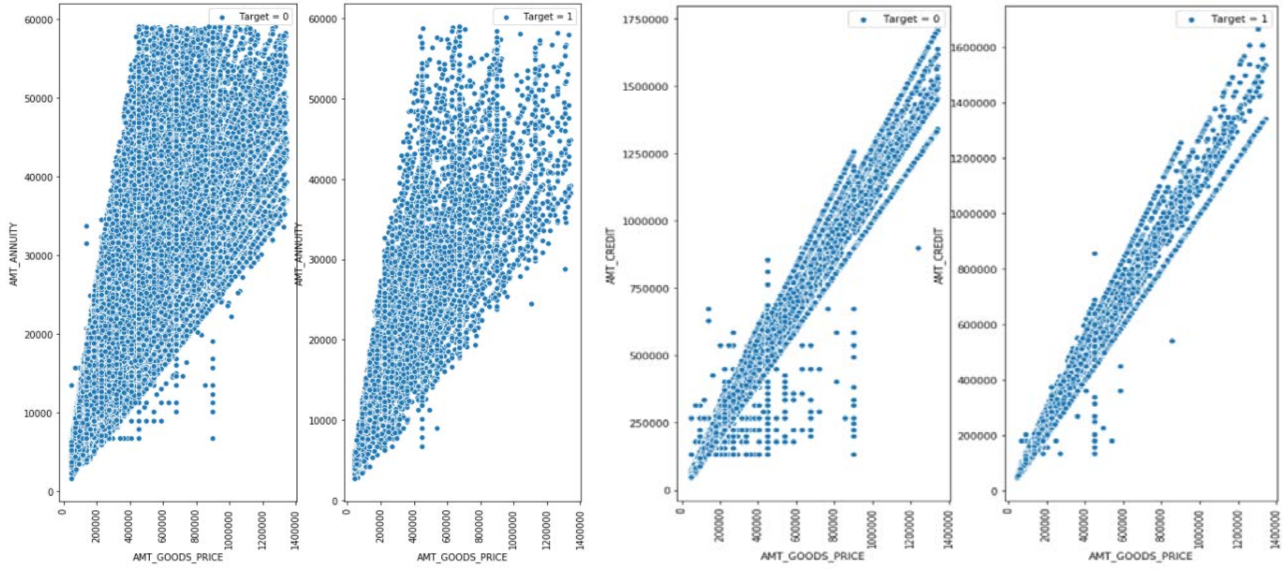
AMT_GOODS_PRICE and AMT_CREDIT –
The correlation between property price and loan amount for non defaulters is 0.9819
but for defaulters it is: 0.9779

AMT_ANNUITY and AMT_CREDIT –
The correlation between AMT_ANNUITY (EMI) and loan amount for non defaulters is 0.7606
but for defaulters it is: 0.7397

AMT_ANNUITY and AMT_GOOD_PRICE –
The correlation between AMT_ANNUITY (EMI) and goods price for non defaulters is 0.7604
but for defaulters it is: 0.7375

AMT_ANNUITY and AMT_INCOME_TOTAL-
The correlation between AMT_INCOME_TOTAL (EMI) and AMT_ANNUITY for non defaulters is 0.4078
but for defaulters it is: 0.3854





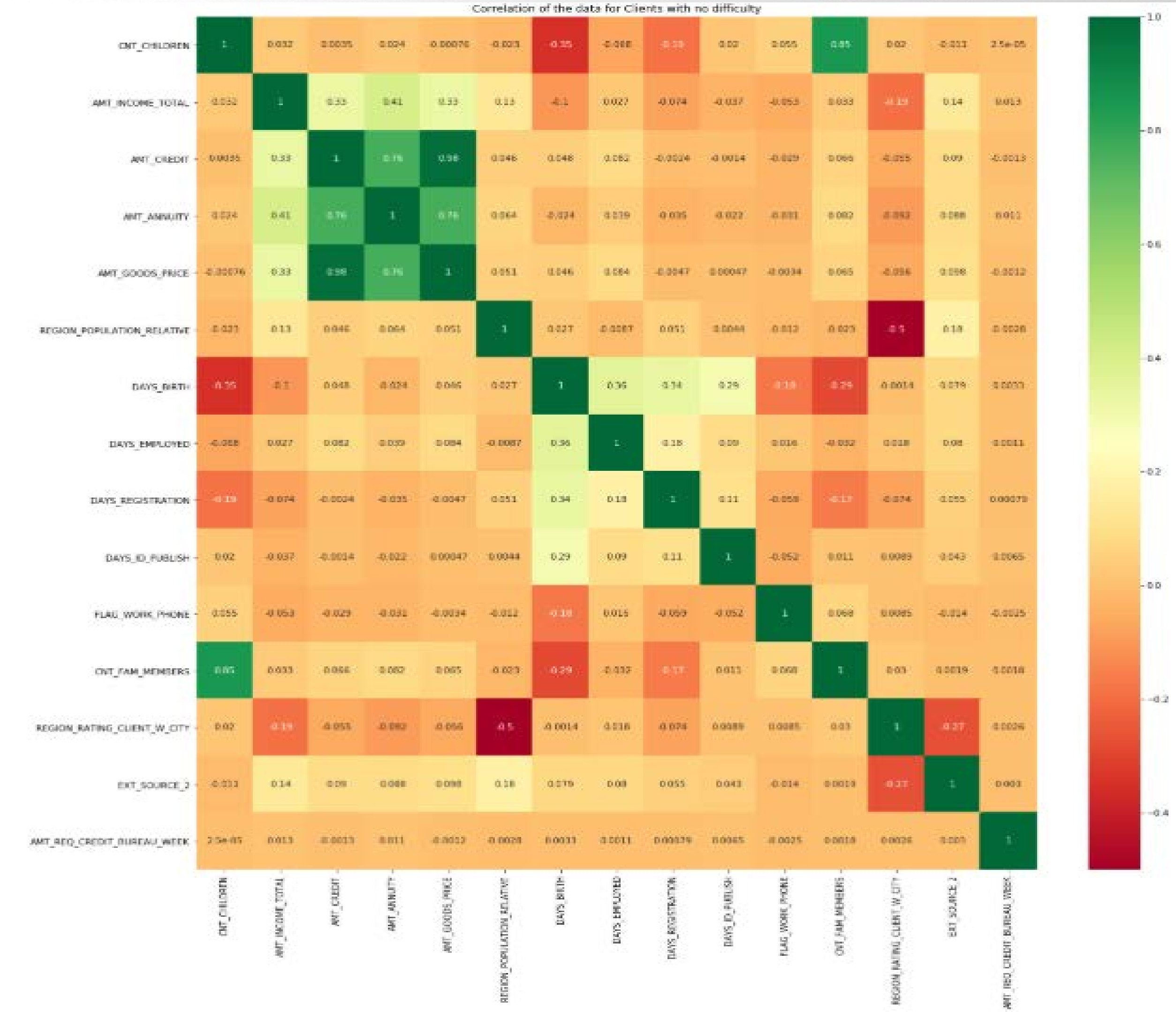
Data Analysis: Correlation Analysis

Top 10 Positively Correlated variables for Target 0:

	VAR1	VAR2	CORR
62	AMT_GOODS_PRICE	AMT_CREDIT	0.981909
165	CNT_FAM_MEMBERS	CNT_CHILDREN	0.853202
47	AMT_ANNUITY	AMT_CREDIT	0.760565
63	AMT_GOODS_PRICE	AMT_ANNUITY	0.760415
46	AMT_ANNUITY	AMT_INCOME_TOTAL	0.407785
111	DAYS_EMPLOYED	DAYS_BIRTH	0.355384
126	DAYS_REGISTRATION	DAYS_BIRTH	0.335582
61	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.330714
31	AMT_CREDIT	AMT_INCOME_TOTAL	0.328006
141	DAYS_ID_PUBLISH	DAYS_BIRTH	0.291728

- **Higher the Good Price for the loans the applicants are applying, higher is the amount credit**
- **Higher their Amount Annuity , higher will be the amount credit**

Data Set: Applicant who find no difficult to repay loan, TARGET: 0



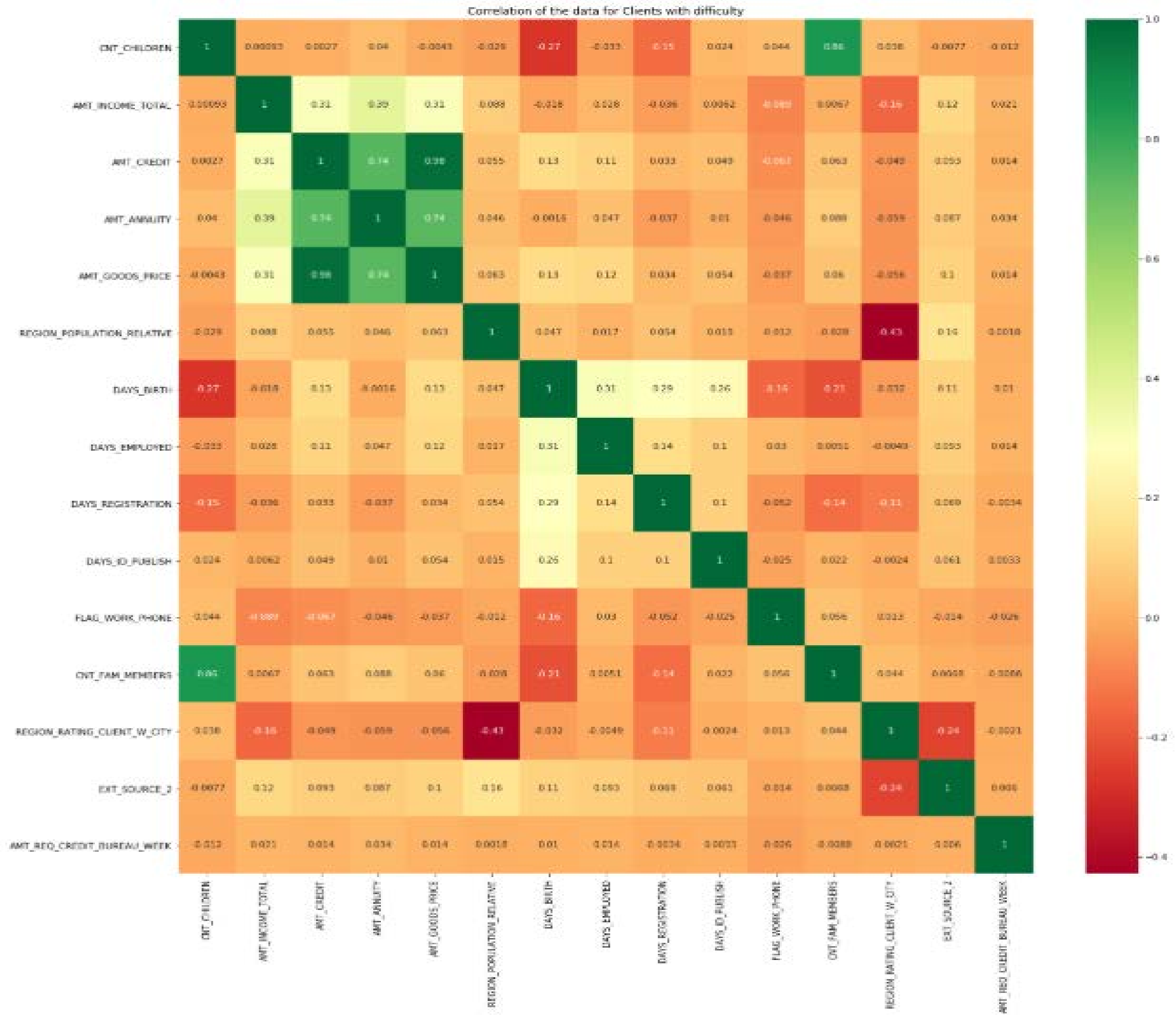
Data Analysis: Correlation Analysis

Top 10 Positively Correlated variables for Target 1:

	VAR1	VAR2	CORR
62	AMT_GOODS_PRICE	AMT_CREDIT	0.977949
165	CNT_FAM_MEMBERS	CNT_CHILDREN	0.858025
47	AMT_ANNUITY	AMT_CREDIT	0.739727
63	AMT_GOODS_PRICE	AMT_ANNUITY	0.737484
46	AMT_ANNUITY	AMT_INCOME_TOTAL	0.385447
111	DAYS_EMPLOYED	DAYS_BIRTH	0.309745
61	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.306330
31	AMT_CREDIT	AMT_INCOME_TOTAL	0.306087
126	DAYS_REGISTRATION	DAYS_BIRTH	0.289750
141	DAYS_ID_PUBLISH	DAYS_BIRTH	0.262175

- Higher the Annuity they have higher Credit
- There is not much correlation between Amount Annuity and income of the individual
- Higher the Goods price they are also having the high Credit and Annuity

Data Set: Applicant who find it difficult to repay loan, TARGET: 1



Historical Data Analysis

Dataset used : Previous_application.csv

Inspected dataset using Shape, Info, Describe and head functions

Shape of the data frame : (1670214, 37)
Total no of columns with missing values > 50% : 4

Dropped the Columns which are populated less than 50 %, as these might not be helpful in giving the right understanding of the data.

Retained only the important columns that are found to be relevant for the analysis.

We have changed the datatypes to suitable type for analysis and also changed negative values to positive.

```
prev_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 69635 entries, 4 to 1670206
Data columns (total 33 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   SK_ID_PREV                   69635 non-null  int64
 1   SK_ID_CURR                   69635 non-null  float64
 2   NAME_CONTRACT_TYPE           69635 non-null  object
 3   AMT_ANNUITY                  69372 non-null  float64
 4   AMT_APPLICATION              69635 non-null  float64
 5   AMT_CREDIT                   69635 non-null  float64
 6   AMT_GOODS_PRICE              69635 non-null  float64
 7   WEEKDAY_APPR_PROCESS_START   69635 non-null  object
 8   HOUR_APPR_PROCESS_START      69635 non-null  int64
 9   FLAG_LAST_APPL_PER_CONTRACT  69635 non-null  object
 10  NFLAG_LAST_APPL_IN_DAY       69635 non-null  int64
 11  NAME_CASH_LOAN_PURPOSE       69635 non-null  object
 12  NAME_CONTRACT_STATUS         69635 non-null  object
 13  DAYS_DECISION                69635 non-null  int64
 14  NAME_PAYMENT_TYPE            69635 non-null  object
 15  CODE_REJECT_REASON           69635 non-null  object
 16  NAME_TYPE_SUITE              42457 non-null  object
 17  NAME_CLIENT_TYPE             69635 non-null  object
 18  NAME_GOODS_CATEGORY          69635 non-null  object
 19  NAME_PORTFOLIO               69635 non-null  object
 20  NAME_PRODUCT_TYPE            69635 non-null  object
 21  CHANNEL_TYPE                 69635 non-null  object
 22  SELLERPLACE_AREA             69635 non-null  int64
 23  NAME_SELLER_INDUSTRY         69635 non-null  object
 24  CNT_PAYMENT                  69372 non-null  float64
 25  NAME_YIELD_GROUP             69635 non-null  object
 26  PRODUCT_COMBINATION          69635 non-null  object
 27  DAYS_FIRST_DRAWING           0 non-null      float64
 28  DAYS_FIRST_DUE               24632 non-null  float64
 29  DAYS_LAST_DUE_1ST_VERSION    24640 non-null  float64
 30  DAYS_LAST_DUE                19596 non-null  float64
 31  DAYS_TERMINATION             19498 non-null  float64
 32  NFLAG_INSURED_ON_APPROVAL    24640 non-null  float64
dtypes: float64(12), int64(5), object(16)
memory usage: 18.1+ MB
```

Merged Data Analysis

Dataset used : inner join (df_cleaned + prev_df)
We have performed inner join on SK_ID_CURR

Inspected dataset using Shape, Info, Describe and Head functions

Shape of the data frame : (53193 , 68)

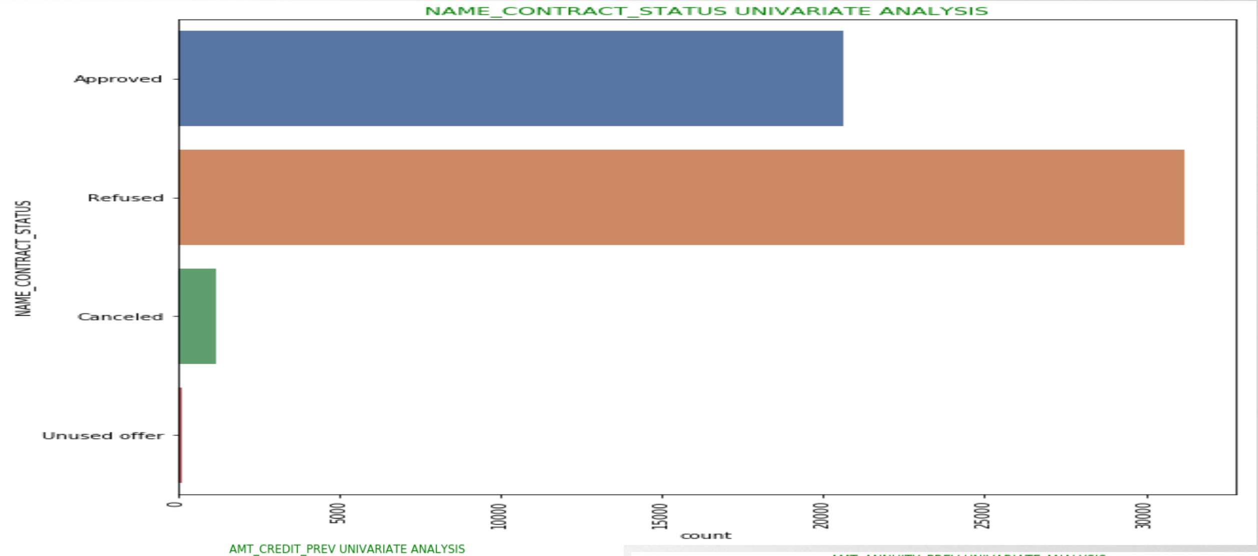
Retained only the important columns that are found to be relevant for the analysis.

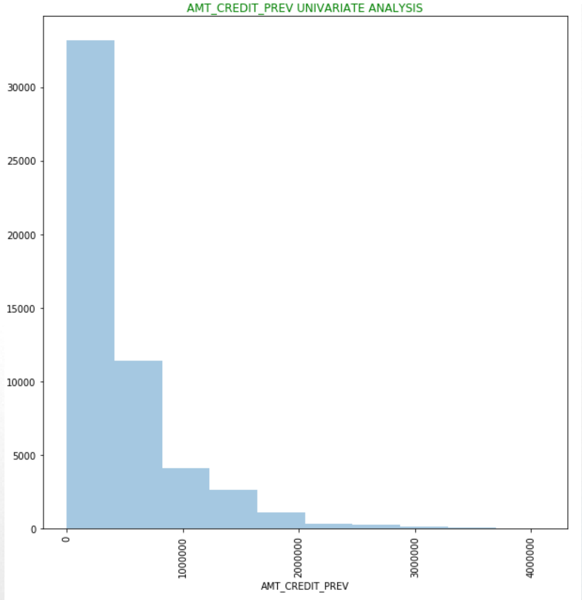
```
mdf.info()
```

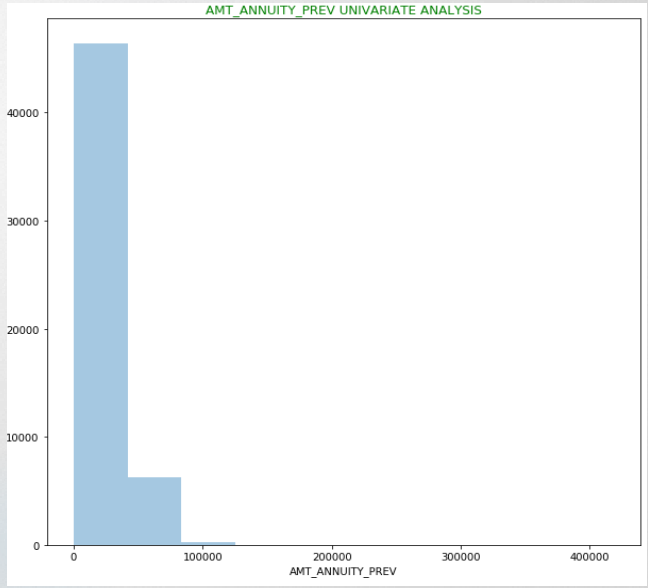
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 53193 entries, 0 to 53192
Data columns (total 63 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   TARGET                      53193 non-null  int64
 1   NAME_CONTRACT_TYPE          53193 non-null  object
 2   CODE_GENDER                 53193 non-null  object
 3   FLAG_OWN_CAR                53193 non-null  object
 4   FLAG_OWN_REALTY             53193 non-null  object
 5   CNT_CHILDREN                53193 non-null  int64
 6   AMT_INCOME_TOTAL            53193 non-null  float64
 7   AMT_CREDIT                  53193 non-null  float64
 8   AMT_ANNUITY                 53193 non-null  float64
 9   AMT_GOODS_PRICE_            53193 non-null  float64
 10  NAME_TYPE_SUITE_            53193 non-null  object
 11  NAME_INCOME_TYPE            53193 non-null  object
 12  NAME_EDUCATION_TYPE         53193 non-null  object
 13  NAME_FAMILY_STATUS          53193 non-null  object
 14  NAME_HOUSING_TYPE           53193 non-null  object
 15  REGION_POPULATION_RELATIVE  53193 non-null  float64
 16  DAYS_BIRTH                  53193 non-null  int64
 17  DAYS_EMPLOYED               45776 non-null  float64
 18  DAYS_REGISTRATION           53193 non-null  int32
 19  DAYS_ID_PUBLISH             53193 non-null  int64
 20  FLAG_MOBIL                  53193 non-null  int64
 21  FLAG_EMP_PHONE              53193 non-null  int64
 22  FLAG_WORK_PHONE             53193 non-null  int64
 23  FLAG_CONT_MOBILE            53193 non-null  int64
 24  FLAG_PHONE                  53193 non-null  int64
 25  FLAG_EMAIL                  53193 non-null  int64
 26  OCCUPATION_TYPE             53193 non-null  object
 27  CNT_FAM_MEMBERS             53193 non-null  int32
 28  REGION_RATING_CLIENT_W_CITY 53193 non-null  int64
 29  ORGANIZATION_TYPE           53193 non-null  object
 30  EXT_SOURCE_2                53193 non-null  float64
 31  AMT_REQ_CREDIT_BUREAU_WEEK  53193 non-null  float64
 32  Age_BIN                     53193 non-null  object
 33  Emp_BIN                     53193 non-null  object
 34  Income_BIN                  53193 non-null  object
 35  SK_ID_PREV                  53193 non-null  int64
 36  NAME_CONTRACT_TYPE_PREV     53193 non-null  object
 37  AMT_ANNUITY_PREV            52980 non-null  float64
 38  AMT_APPLICATION             53193 non-null  float64
 39  AMT_CREDIT_PREV             53193 non-null  float64
 40  AMT_GOODS_PRICEx            53193 non-null  float64
 41  NAME_CASH_LOAN_PURPOSE      53193 non-null  object
 42  NAME_CONTRACT_STATUS        53193 non-null  object
 43  DAYS_DECISION               53193 non-null  int64
 44  NAME_PAYMENT_TYPE           53193 non-null  object
 45  CODE_REJECT_REASON          53193 non-null  object
 46  NAME_TYPE_SUITEx            31804 non-null  object
 47  NAME_CLIENT_TYPE            53193 non-null  object
 48  NAME_GOODS_CATEGORY         53193 non-null  object
 49  NAME_PORTFOLIO              53193 non-null  object
 50  NAME_PRODUCT_TYPE           53193 non-null  object
 51  CHANNEL_TYPE                53193 non-null  object
 52  SELLERPLACE_AREA            53193 non-null  int64
 53  NAME_SELLER_INDUSTRY        53193 non-null  object
 54  CNT_PAYMENT                 52980 non-null  float64
 55  NAME_YIELD_GROUP            53193 non-null  object
 56  PRODUCT_COMBINATION         53193 non-null  object
 57  DAYS_FIRST_DRAWING          0 non-null      float64
 58  DAYS_FIRST_DUE              18914 non-null  float64
 59  DAYS_LAST_DUE_1ST_VERSION   18920 non-null  float64
 60  DAYS_LAST_DUE               14767 non-null  float64
 61  DAYS_TERMINATION            14685 non-null  float64
 62  NFLAG_INSURED_ON_APPROVAL   18920 non-null  float64
dtypes: float64(19), int32(2), int64(14), object(28)
memory usage: 28.1+ MB
```

Univariate Analysis on merged data

- The Amount Annuity is highly distributed below 80,000
- We observe that number of refused applicant is higher in Name contract status of previous application dataset
- The Amount credit previous data is highly distributed below 10 lakhs
- Most rejection of loans came from purpose 'repairs'.
- For education purposes we have equal number of approves and rejection
- Paying other loans and buying a new car is having significant higher rejection than approves
- Here the proportion of Name contract distribution between M and F values is the same
- Here in working income source we have it has major category and have higher refused applicants. when compared to other income sources

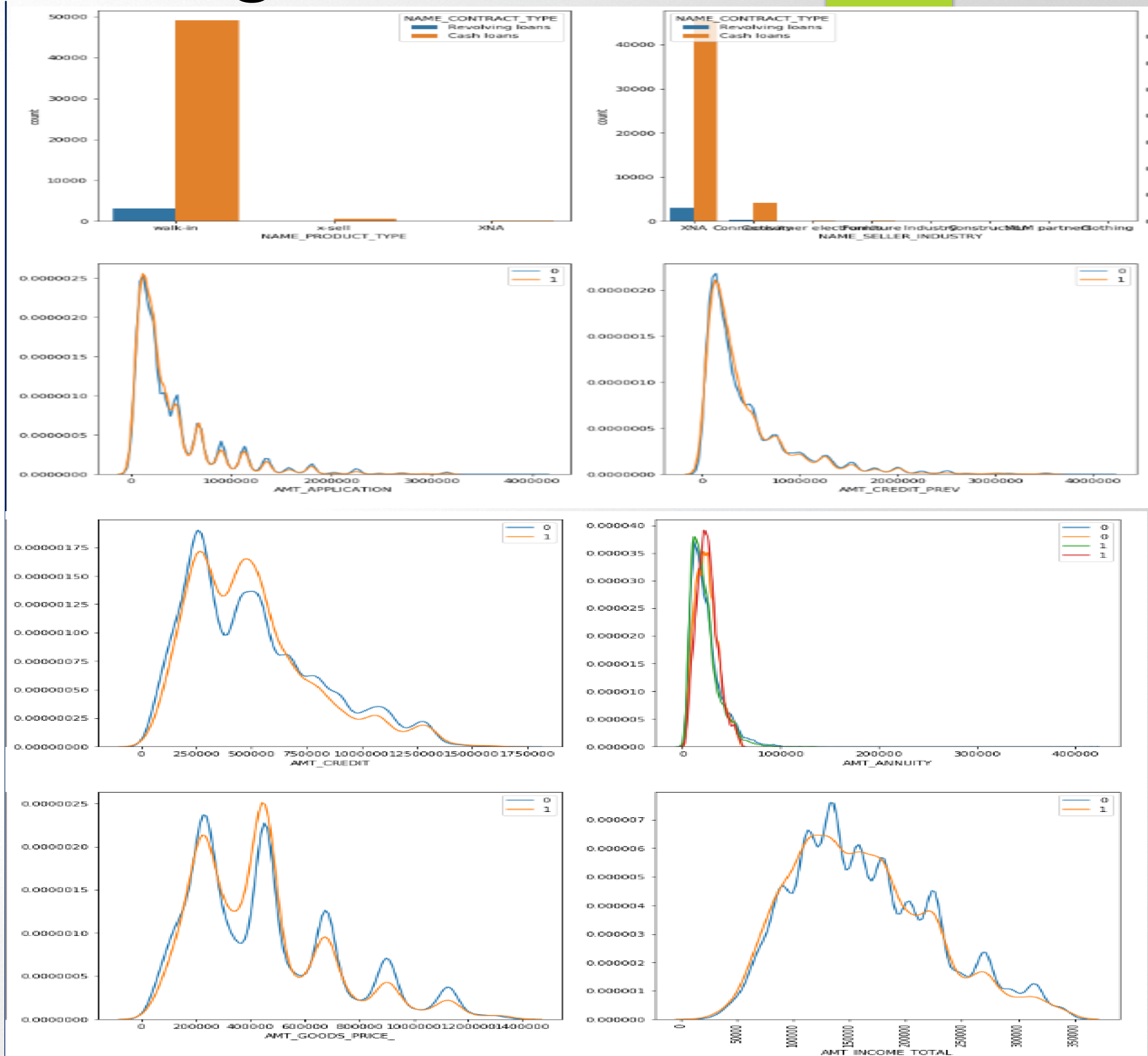






Bivariate Analysis on merged data

- The Applicants in Name Product type - Xsell have only cash loans and no revolving loans
- The Amount annuity in range 80,000 - 1,00,000 have no difficulty in repaying the loan
- Applicants who are having the Annuity below the 80k in the history are most likely to have problem in repaying the loan, It could be possibly they might be taking the new loan for the annuity itself
- Applicants who has taken loan amount in range 30lakhs to 40lakhs previously are having no trouble paying the loan now or the data is insufficient to make an inference
- If they have a credit history less than 5 Lakhs, the count of people repaying the amount is more than who can’t.
- For the applicants who have taken loan less than 500K work of goods price might face problem in repaying the loan now
- Applicants who has amount income total between 1lakh to 3.5lakh don't face difficulty in repaying the loan



Recommendations

The dataset is highly imbalanced with 8.35% data for Loan Defaulters and remaining 91.65% data for non-defaulters.
In the application dataset: The top 10 positive correlation between numerical variables, is consistent across both Defaulters and non-defaulters

- Bank should focus more on people with annual income between 1lakh - 3.5lakh as they don’t face any difficulty in repaying the loan
- Applicants with higher Income could have reactively higher credit and annuity but chances of them defaulting / late payment is less and should be focused more by banks when approving loan
- Banks should focus less on income type ‘Working’ as they are having most number of unsuccessful payments.
- Most rejection of loans came from purpose 'repairs', so bank should focus less on this section
- The bank should focus less on younger customers as they have more number of defaulters


THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
DANKSCHEEN
JUSPAXAR
BİYAN
SHUKRIA
TINGKI
YAQHANYELAY
TASHAKKUR ATU
SUKSAMA
EKHMET
MEHRBANI
GRAZIE
MAAKE
KOMAPSUMNIDA
PALDIES
BOLZÏN
MERCI
GOZAIMASHITA
EFCHARISTO
SPASIBO
DENKAUJA
NENACHALHYA
SPASSIBO
SNACHALHUYA
NUHUN
CHALTU
WABEEJA
MAITEKA
HUI
GUI
HATUR
EKOJU
SIKOMO
MAKETAI
MINMONCHAR
MERASTAWHY
GAEJTHO
AGUYJE
FAKAAUE
LAH
TAVTAPUCH
MEDAWAGSE
ATTO
MERSI
UNALCHÉESH
YUSPAGARATAM