EDA Credit

Manu Madhavan(manucet439@Gmail.com)

Problem Statement

- Loan providers struggle with approving loans for people with little or no credit history.
- Loan approval is based on the applicant's risk profile.
- Two types of risk are associated with loan approval decisions:
    - Denying a loan to a reliable applicant results in lost business for the company.
    - Approving a loan to an unreliable applicant may lead to financial loss for the company

Objective

- This case study aims to identify patterns indicating if a client may have difficulty paying instalments.
- These patterns will be used to inform actions such as denying the loan, reducing loan amounts, or lending at a higher interest rate to risky applicants.
- The goal is to ensure that consumers who can repay the loan are not rejected.
- Applications from applicants who are not capable of paying back the loan should be rejected.

Application Data Analysis

Missing value identification and Imputation

Identification

Missing values are identified in both the data set using below sample command.
Columns with >30 % Missing values are dropped.

```
# Calculate the percentage of missing values in each column
missing_percentages = inp1.isnull().sum() / len(inp1)
```

```
missing_percentages.sort_values(ascending=False).head(60)*100
```

```
COMMONAREA_MEDI             69.872297
COMMONAREA_AVG              69.872297
COMMONAREA_MODE             69.872297
NONLIVINGAPARTMENTS_MODE    69.432963
```

```
((inp2.isnull().sum()/len(inp2))*100).sort_values(as
```

```
EXT_SOURCE_3                  19.825307
AMT_REQ_CREDIT_BUREAU_YEAR    13.501631
AMT_REQ_CREDIT_BUREAU_QRT     13.501631
AMT_REQ_CREDIT_BUREAU_MON     13.501631
AMT_REQ_CREDIT_BUREAU_WEEK    13.501631
AMT_REQ_CREDIT_BUREAU_DAY     13.501631
AMT_REQ_CREDIT_BUREAU_HOUR    13.501631
NAME_TYPE_SUITE                0.420148
DEF_60_CNT_SOCIAL_CIRCLE       0.332021
OBS_30_CNT_SOCIAL_CIRCLE       0.332021
DEF_30_CNT_SOCIAL_CIRCLE       0.332021
OBS_60_CNT_SOCIAL_CIRCLE       0.332021
EXT_SOURCE_2                   0.214626
AMT_GOODS_PRICE                0.090403
AMT_ANNUITY                    0.003902
CNT_FAM_MEMBERS                0.000650
DAYS_LAST_PHONE_CHANGE         0.000325
```

Imputation-1

Missing values are filled with existing value

EXT_SOURCE_3- missing value filled with Value of EXT_SOURCE 2 and wise versa

Assumption 1:Min ,Max,Median value or almost similar range
Assumption 2:Its is user feedback from external source

Imputation- 2

Missing values are filled with Mode

For categorical Values missing value are imputed with Highest occurring values.

Imputation- 3-Drop missing value

For Numerical value, best approach in this data set for marked column is to drop rows, as the % of missing value is very less

Assumption: Standard deviation is very high so taking average is not good idea

While checking on categorical column values,its identified that Gender and Organization Type have XNA. In Gender XNA is very less, hence rows dropped but in Organization Type XNA is 18% hence not taken this column for any analysis

```
inp2.CODE_GENDER.value_counts()
```

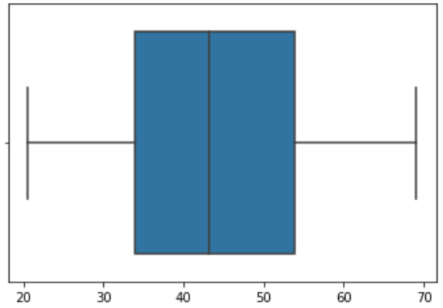
```
F        202448
M        105059
XNA           4
Name: CODE_GENDER, dtype: int64
```

Transformation

DAYS_BIRTH and DAYS_EMPLOYED are in negative which is changed to positive value and converted to year

```
inp2.DAYS_BIRTH=inp2.DAYS_BIRTH.apply(lambda x:abs(x)/365.25)
```

```
sns.boxplot(inp2.DAYS_BIRTH)
plt.show()
```



DAYS_BIRTH and DAYS_EMPLOYED Further converted to Buckets during the analysis

```
#Categorizing the Age group and Year Employed in to different bukets.

inp2_Target1['Age_Group']=pd.cut(inp2_Target1.DAYS_BIRTH[:5],[0, 30, 40, 50, 60, 9999], labels= ["<30","30-40","40-50","50-60",
```

```
inp2_Target0['Age_Group']=pd.cut(inp2_Target0.DAYS_BIRTH[:5],[0, 30, 40, 50, 60, 9999], labels= ["<30","30-40","40-50","50-60",
```

```
inp2_Target1['Employed_Year_Group']=pd.cut(inp2_Target1.DAYS_EMPLOYED[:5],[0, 2, 5, 10, 15, 9999], labels= ["<2","2-5","5-10","
```

```
inp2_Target0['Employed_Year_Group']=pd.cut(inp2_Target0.DAYS_EMPLOYED[:5],[0, 2, 5, 10, 15, 9999], labels= ["<2","2-5","5-10","
```

Data Imbalance Check

```
inp2.TARGET.value_counts()
```

```
0    282682
1     24825
Name: TARGET, dtype: int64
```

```
#Majority of the data is target 0
#Calculating imbalence %
round(len(inp2_Target1)/len(inp2_Target0),2)*100
```

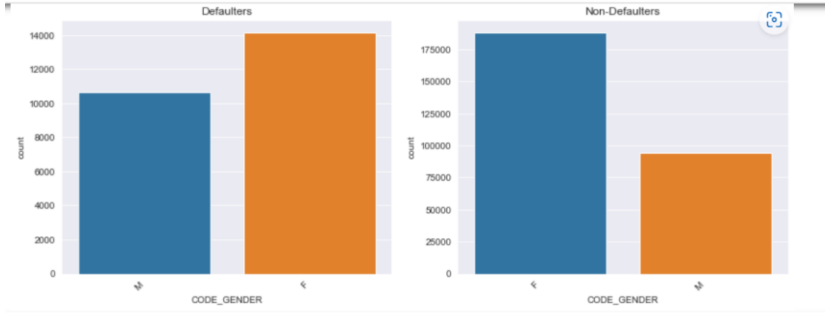
```
9.0
```

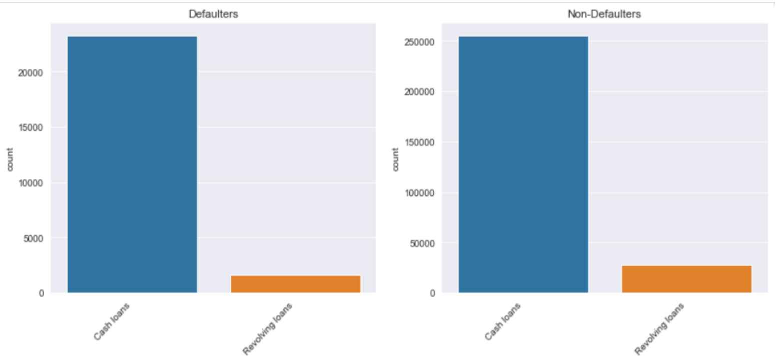
Univariant Analysis on Current Application

Categorical Unordered Univariant Analysis

Observation
1)Contract Type
Defaulters: Cash Loans customer are High in Number compared to revolving loan
Non-Defaulters: Same here, Cash Loans customer are High in Number
2)Gender:
Defaulters: Females are higher
Non-Defaulters: Females are higher
3)Is customer own car:
Defaulters: Most of them not owning a car
Non-Defaulters: Same here ,Most of them not owning a car
4)Own Any reality:
Defaulters: Most of them not owning a reality
Non-Defaulters: Same here
5)Income type:
Defaulters: Majority are working class
Non-Defaulters: Here as well Majority are working class

6)Education type:
Defaulters: Majority are Secondory/Sceondory Special
Non-Defaulters: Majority are Secondory/Sceondory Special
7)Family Status:
Defaulters :Majority are Married
Non-Defaulters: Majority are Married
8)Housing type:
Defaulters: Majority have house/Apartment
Non-Defaulters :Majority have house/Apartment





<u>**Conclusion**</u>
Pattern is same for Defaulters and Non-Defaulters
Female Applicant are higher in both case.
Cash Loan is Higher in Both case.
Most of them not owning a car. Majority are Married having house/Apartment and education secondary special

Univariant Analysis for Numerical Data

- AMT_CREDIT-Same pattern for Defaulter and Non Defaulter,Higher application is lesser amount(Lesser than 1000000)
- DAYS_BIRTH-There is difference in pattern.

Defaulter Density is higher at Age 25 to 30

Non- Defaulter - Density is higher at Age 35 to 40

- DAYS_EMPLOYED-Same Pattern

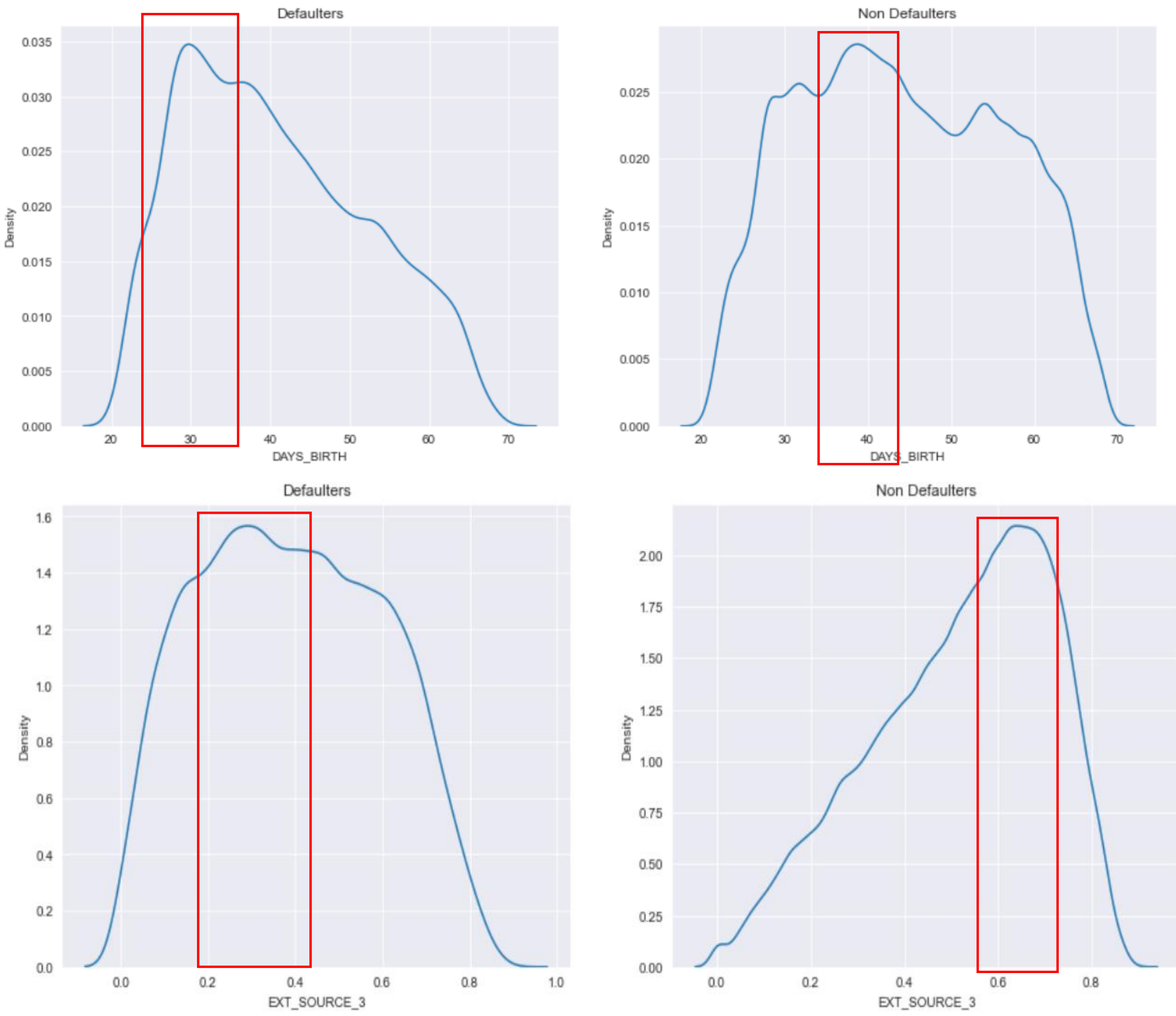
- EXT_SOURCE_3

Defaulter-Defaulter rating is comparatively less(less than .6)

Non Defaulter Rating Density is .6 to 7

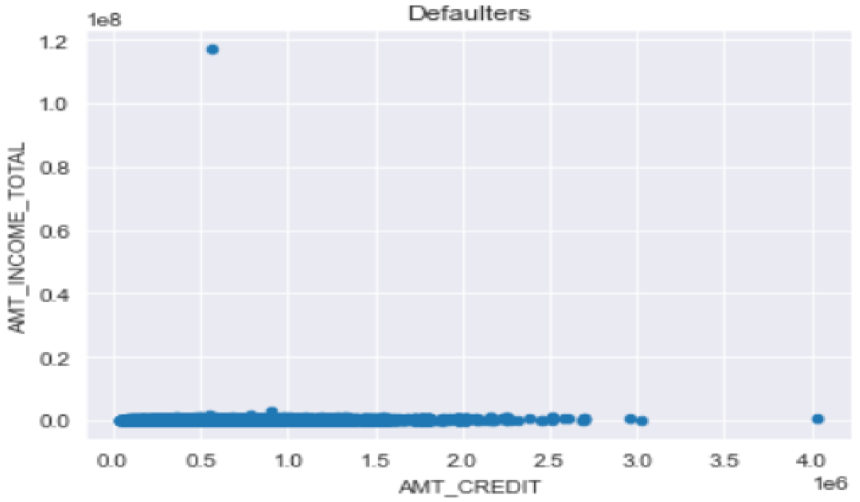
<u>Conclusion</u>

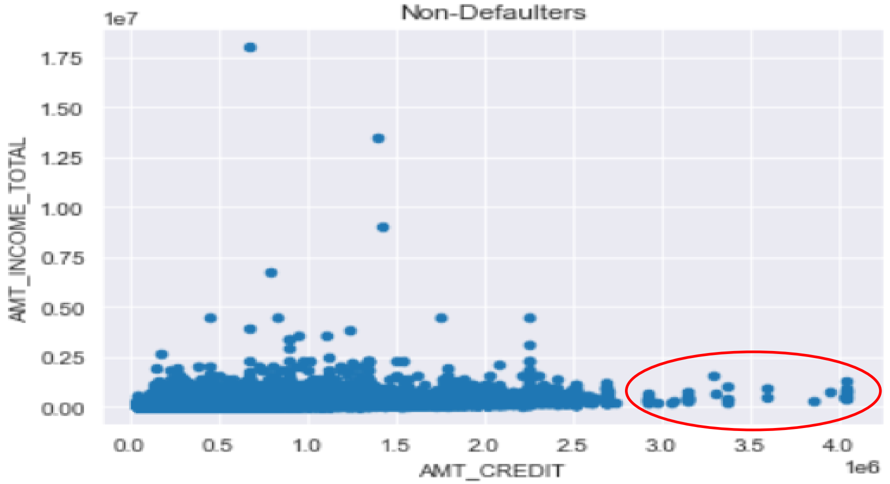
Bank Should **approve loan** of Applicant between **35 to 40** Year of age
And should **reduce approval of age 25 to 30**
Bank Should **approve loan of Applicant with External rating >.6**



BI Variant Analysis(Numeric Numeric)

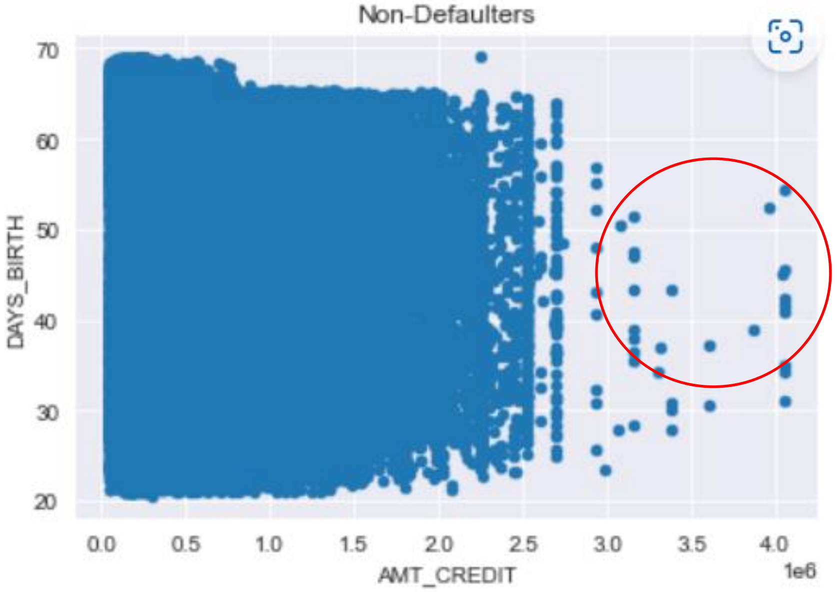
Loan Credited Vs Total Income of Individual (AMT_CREDIT VS AMT_INCOME_TOTAL)





Bank Should provide loan to applicant applying for more than 3000000.

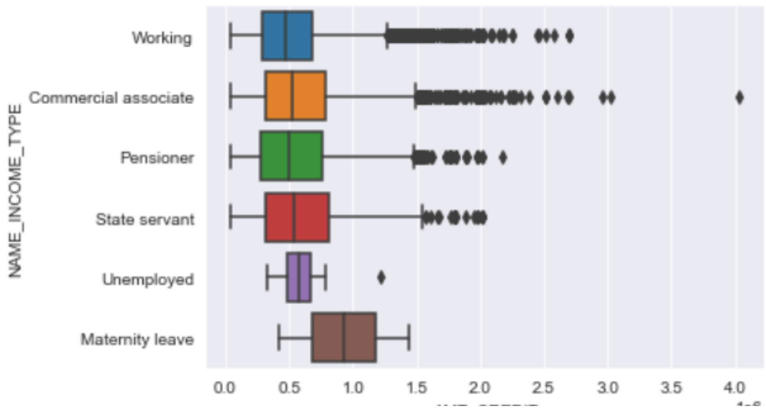
AGE VS AMT Credit

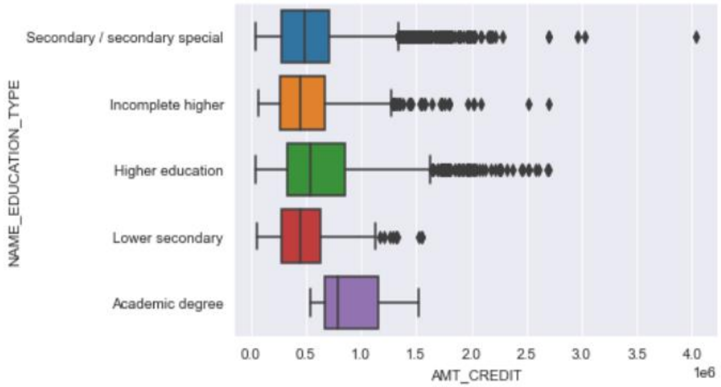


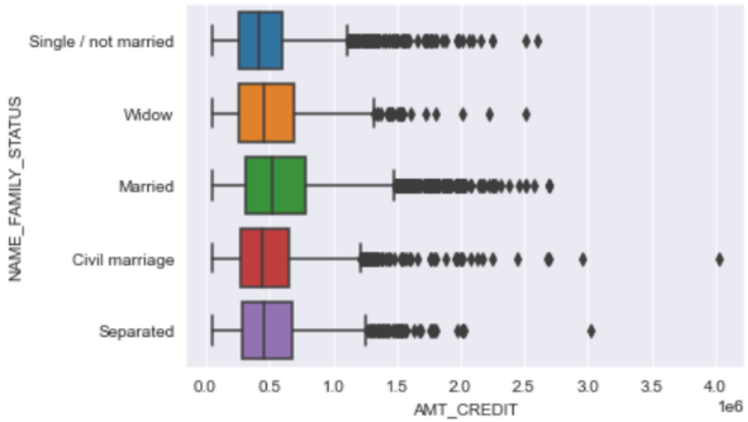
Bank Should not reject any Application from Applicant between Age 35 to 50

BI Variant Analysis(Numeric –Categoric)

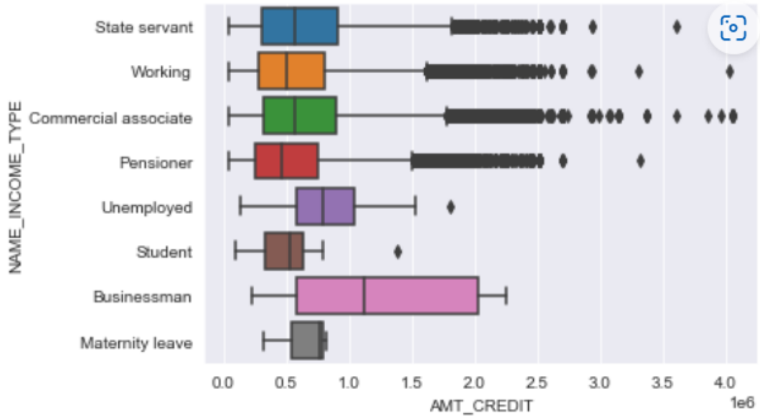
Defaulters

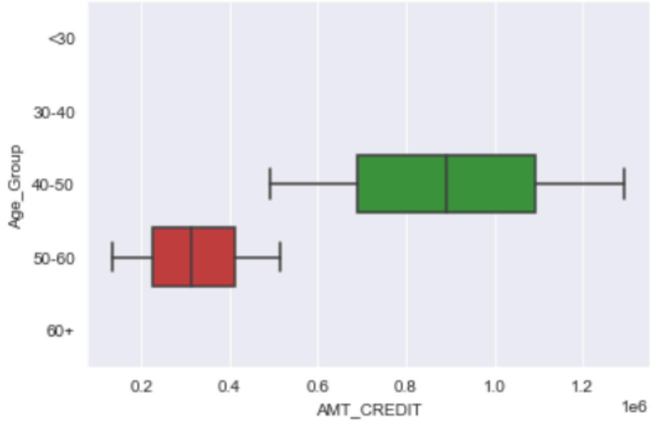


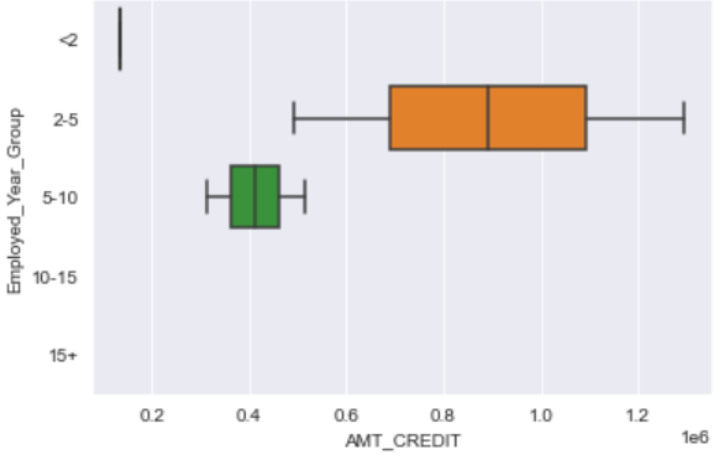




Non-Defaulters







Take Away:

Bank Should be cautious while giving Loan to Applicant who is Married ,Under Maternity leave and having academic degree.

Bank Should Entertain loan for Business Man ,between 40 to 50 Years of age.Also Employee group with 2-5 year of experience

Previous Application Data Analysis

Analysis on Previous Application

```
In [131]: Previous_inp1.isnull().sum()
          ((Previous_inp1.isnull().sum()/len(Previous_inp1))*100).sort_values(ascending=False)

Out[131]: RATE_INTEREST_PRIVILEGED      99.643698
          RATE_INTEREST_PRIMARY         99.643698
          AMT_DOWN_PAYMENT              53.636480
          RATE_DOWN_PAYMENT             53.636480
          NAME_TYPE_SUITE               49.119754
          NFLAG_INSURED_ON_APPROVAL     40.298129
          DAYS_TERMINATION              40.298129
          DAYS_LAST_DUE                 40.298129
          DAYS_LAST_DUE_1ST_VERSION     40.298129
          DAYS_FIRST_DUE                40.298129
          DAYS_FIRST_DRAWING            40.298129
          AMT_GOODS_PRICE               23.081773
          AMT_ANNUITY                   22.286665
          CNT_PAYMENT                   22.286366
          PRODUCT_COMBINATION            0.020716
          AMT_CREDIT                     0.000060
```

1.Columns with >30 % missing are dropped from Analysis

2.Irrelevant columns are dropped from Analysis

```
In [136]: #Drop Some of the irrelavent columns
          irrelavent_col=['NFLAG_LAST_APPL_IN_DAY','NFLAG_LAST_APPL_IN_DAY',
          'FLAG_LAST_APPL_PER_CONTRACT',
          'HOUR_APPR_PROCESS_START',
          'WEEKDAY_APPR_PROCESS_START']

          Previous_inp2 = Previous_inp2.drop(irrelavent_col,axis=1)
```

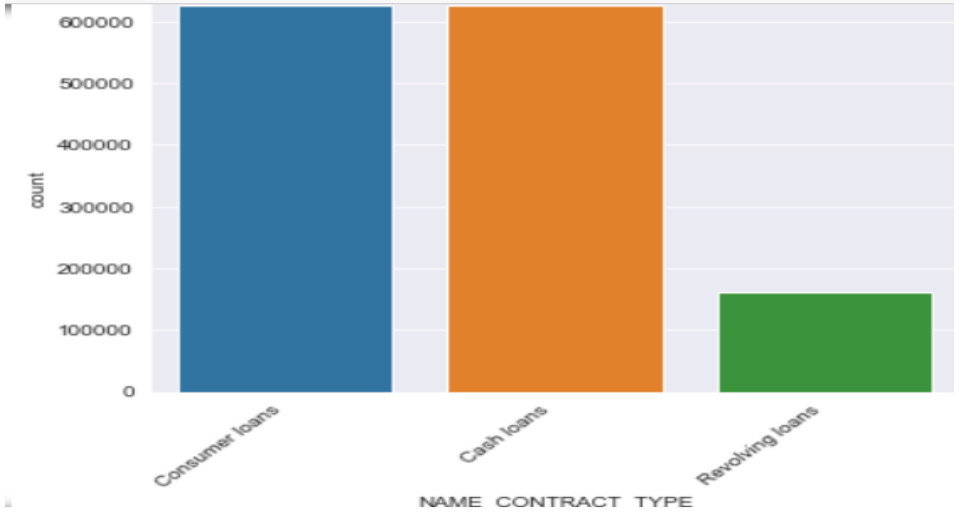
3.Column with XNA & XAP are converted to null

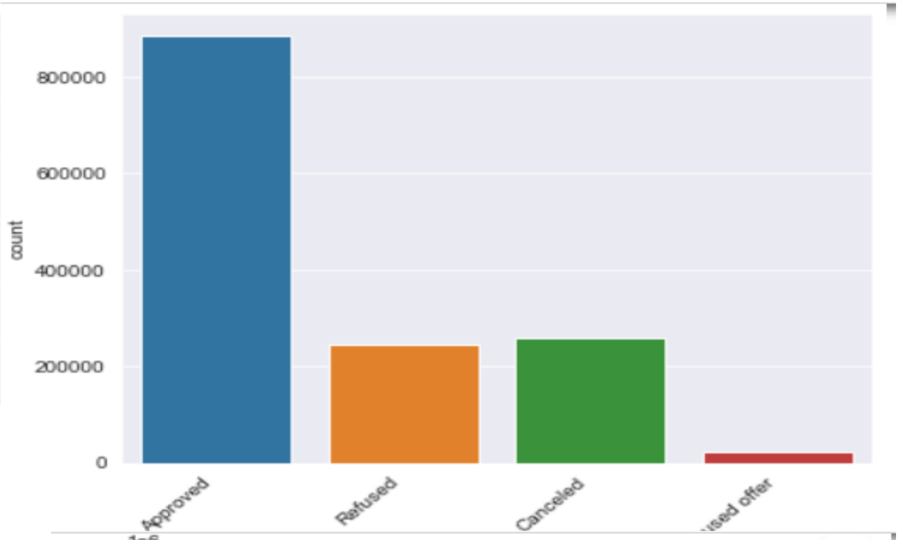
4.For #AMT ANNUITY ,AMT_Goods_price,Name portfolio and CNT_Payment have more than 20 % of missing values.
We may drop this rows from the analysis.

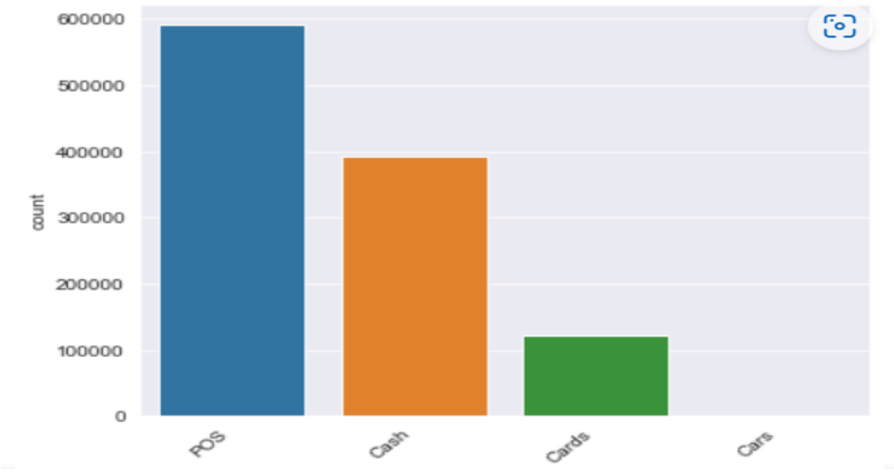
5.Days Decision converted to Month Decision

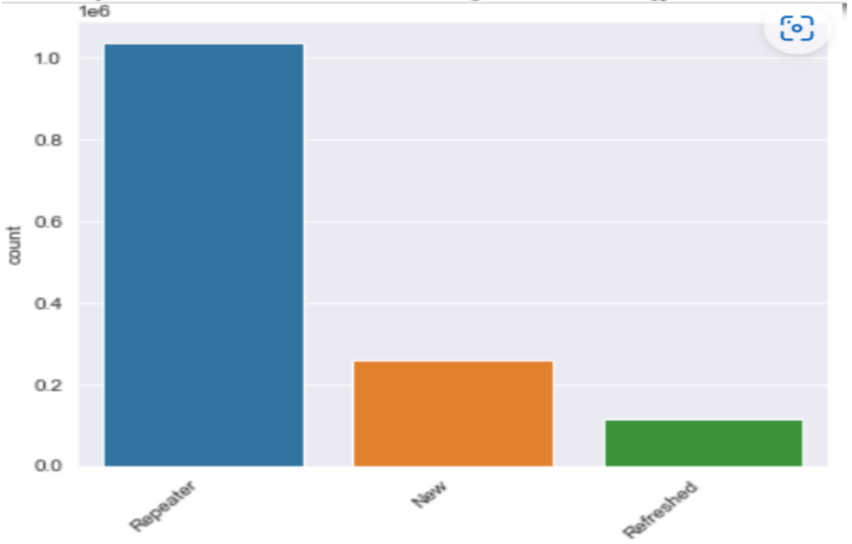
```
Previous_inp2['Month_Decision']=Previous_inp2.DAYS_DECISION.apply(lambda x:abs(x)/30)
```

Univariant Analysis



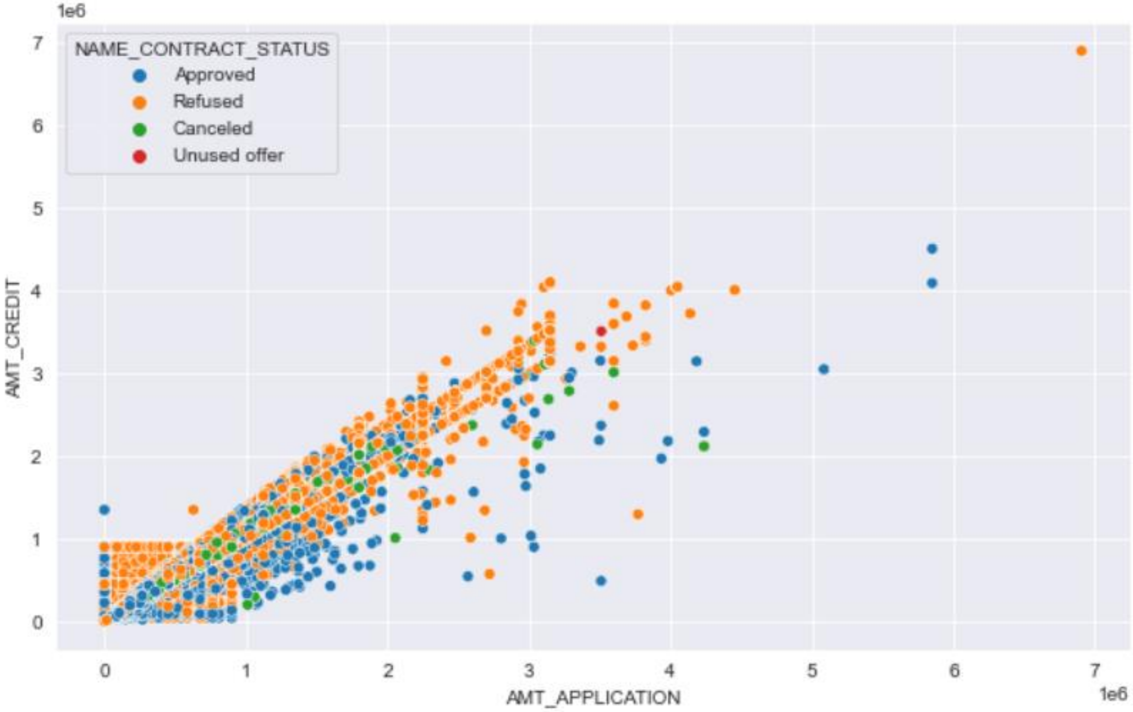


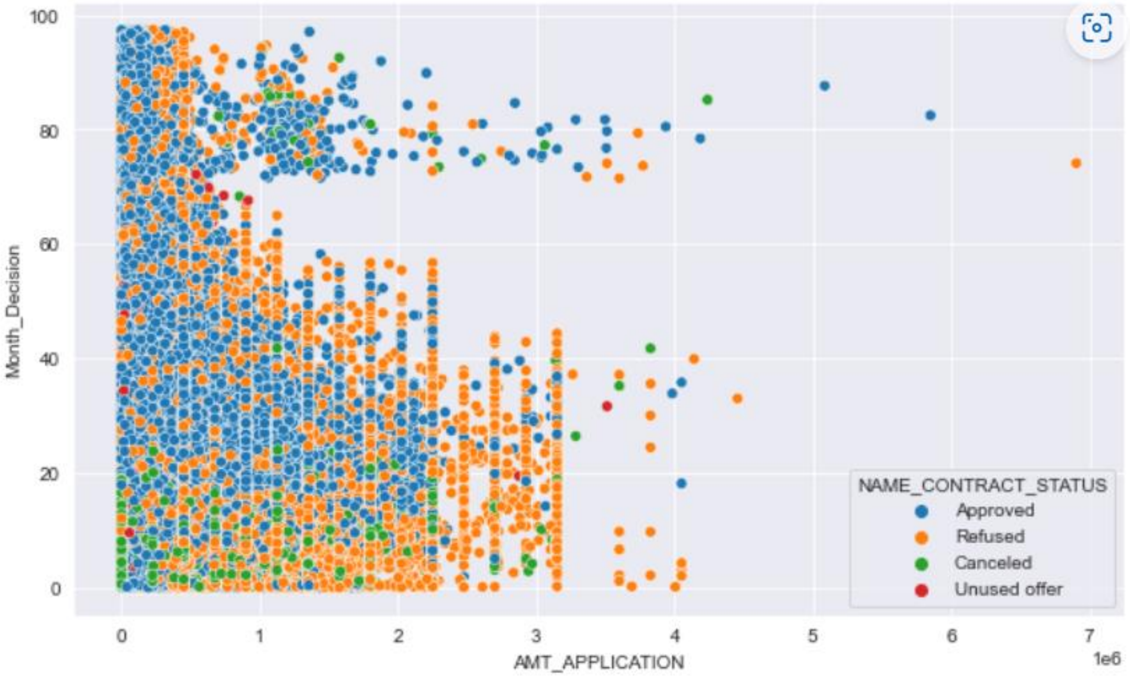




- Most of the previous applications are Consumer loans and Cash loans
- Most of the previous applications are approved
- Application for POS is higher then comes Cash
- Most of the applicant applied for loan multiple times

Bi Variant Analysis





More application are around lesser amount, time taken for taking decision on lesser comparatively higher

Merging of Current and Previous Application

Merging of Current and previous application

Step 1-Few Useful columns selected

```
#Selct columns to merge  from current application

Current_app_cols_to_merge=['SK_ID_CURR', 'TARGET','CODE_GENDER','AMT_INCOME_TOTAL','DAYS_BIRTH',
       'DAYS_EMPLOYED','NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE','NAME_FAMILY_STATUS',
       'NAME_HOUSING_TYPE','EXT_SOURCE_2',
       'EXT_SOURCE_3']
```

Step 2:left merge on previous application on common key SK_ID_Curr.

Step 3:Some of the Application details are not present in Current application,so multiple rows with NAN Generated.

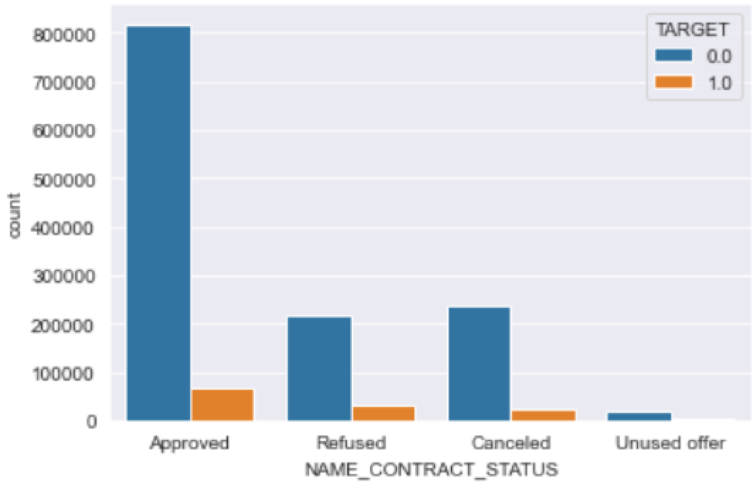
Step 4:Target with 'NAN' rows are dropped

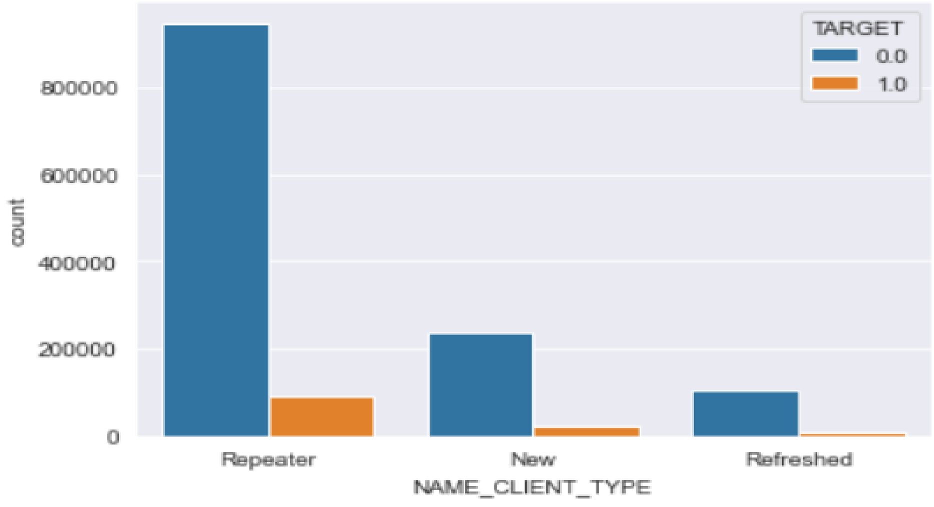
Step 5:Data imbalance check(Defaulter is very less)

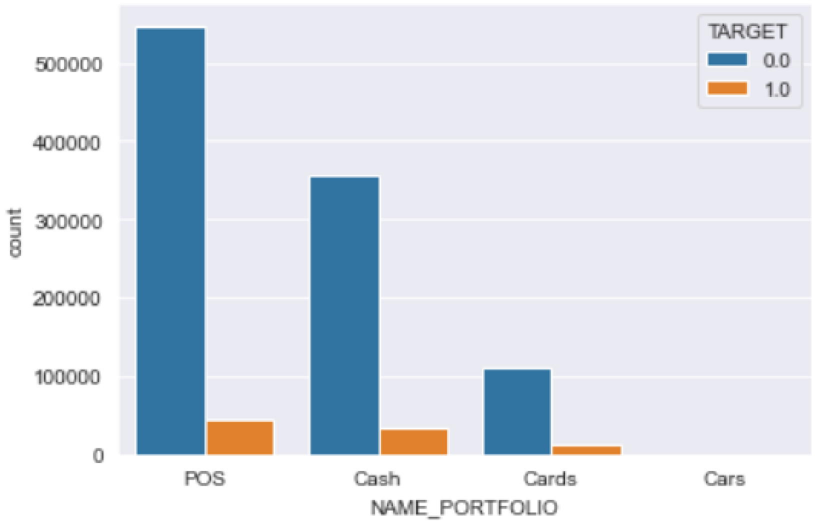
```
Merged_df.TARGET.value_counts()
```

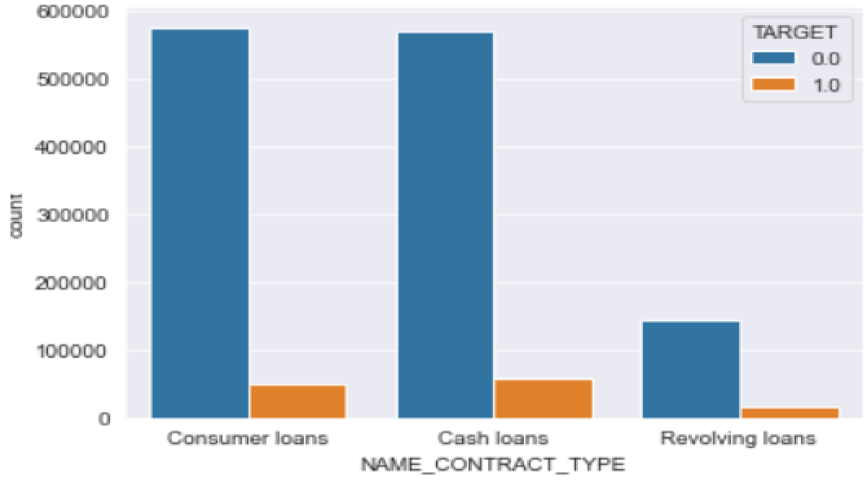
```
0.0    1291286
1.0     122360
Name: TARGET, dtype: int64
```

Univariant Analysis



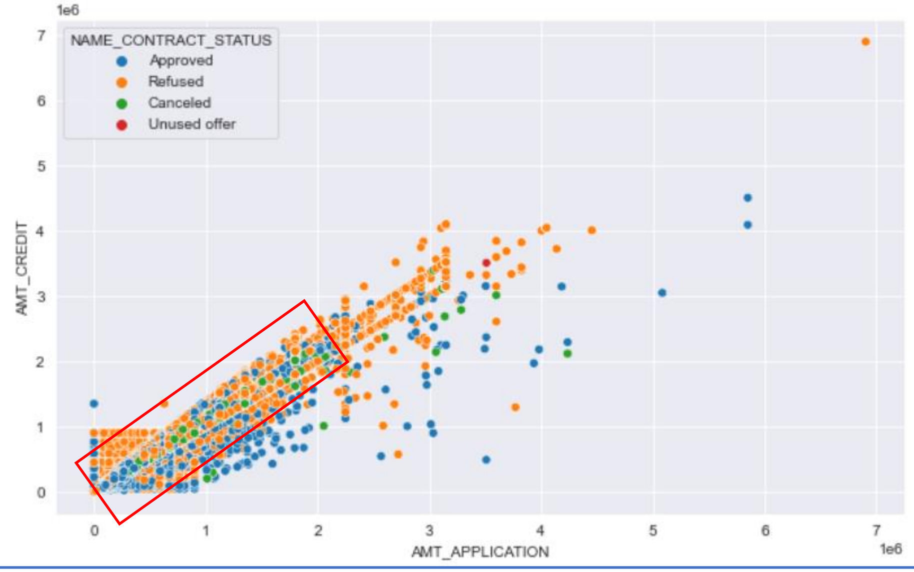


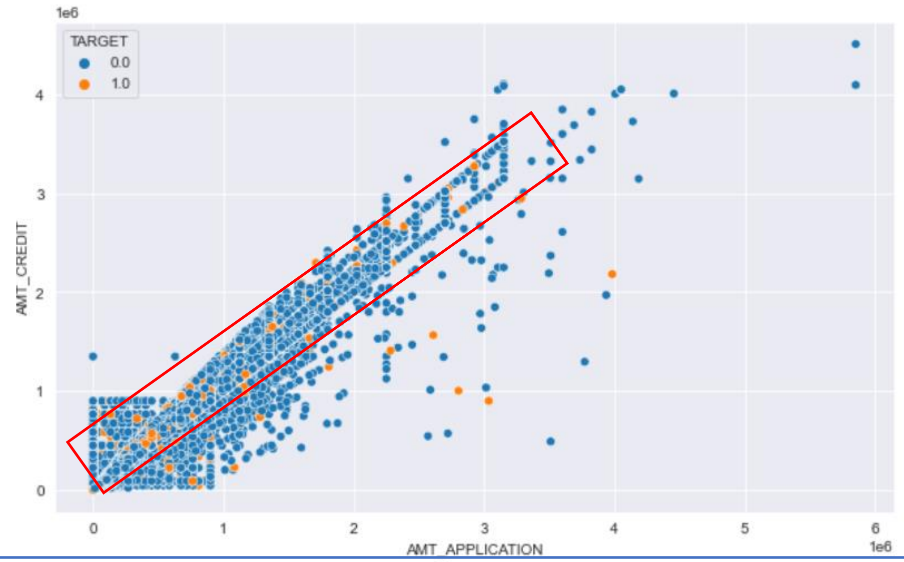




- ➢ There are Non Defaulters in refused category, Bank Should review such customer and provide loan.
- ➢ There are few Defaulters in Approved category
- ➢ Repeated Loan Applicant are less likely to Default.

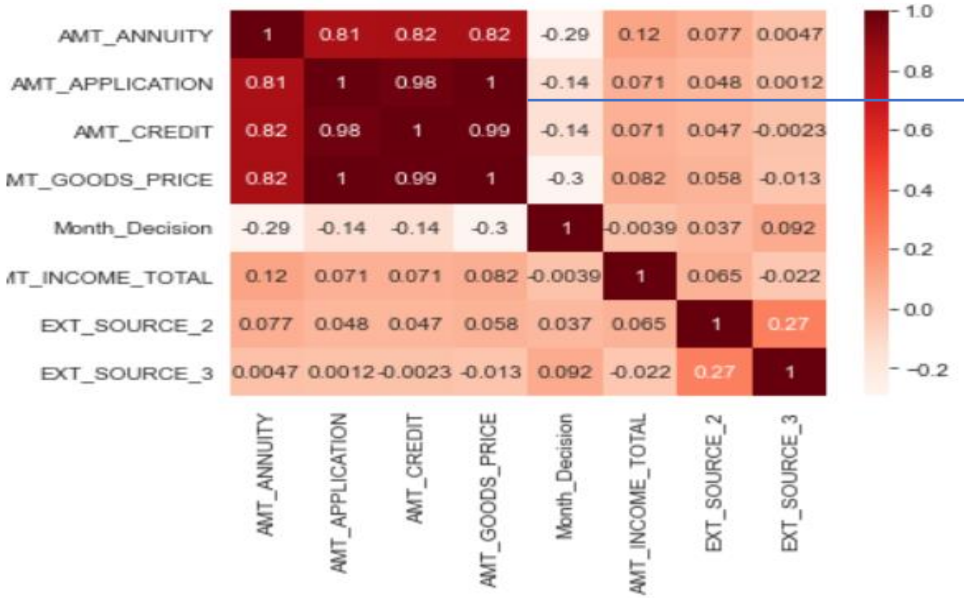
BI Variant Analysis





Bank should review its Criteria of rejection for lesser amount applicant as most of them are not defaulter

Previous Applicant with less external score default rate is comparatively high



- AMT Credit and Amount Goods price have very high correlation
- AMT Application and AMT credit also have high correlation

Conclusion

<u>Conclusion</u>

- Bank Should **approve loan** of Applicant between **35 to 50** Year of age.
- Bank Should **Entertain loan for Business Man ,between 40 to 50 Years of age.**

  Also Employee group with 2-5 year of experience
- Bank Should **approve loan of Applicant with External rating >.6**
- Bank Should provide loan to applicant applying for more than 3000000.
- Bank Should be **cautious while giving Loan to Applicant who is Married ,Under Maternity leave** and having academic degree.
- And should **reduce approval of age 25 to 30.**
- More application are around lesser amount, time taken for taking decision on lesser comparatively higher,bank should improve it