EDA- Case Study

Sapan Kumar

Sourabh Shrivas

Table of Content

Execution Methodology

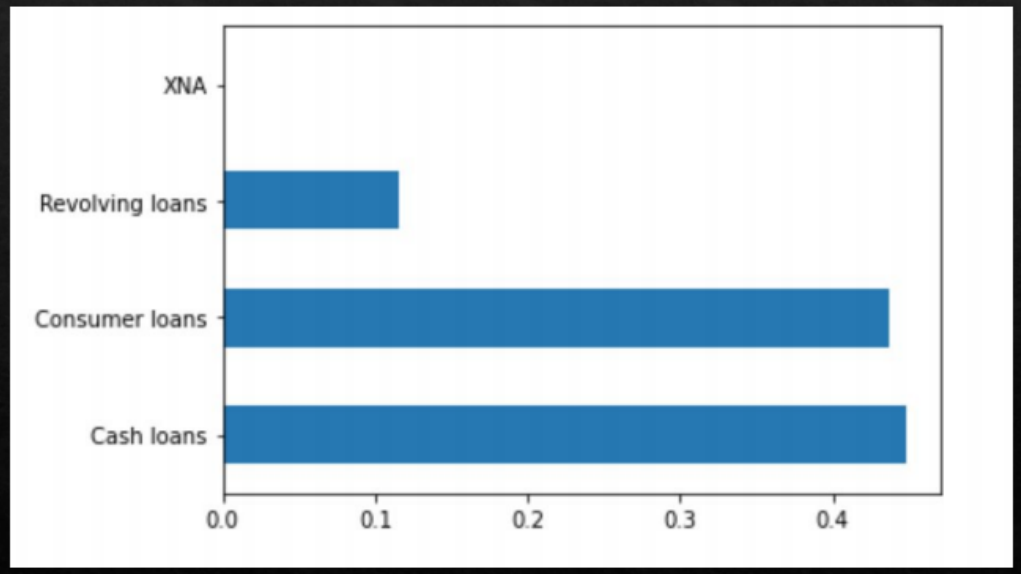
- Data cleaning and data analysis have been done separately for both “Previous data” and “Application data” in two separate Python notebooks
- The insights and inferences were exchanged frequently to make sure we follow similar approach for both the data frames
- All basic steps were followed for Data cleaning, reporting, standardization and further analysis
- Key findings and inferences were summarized in the last page of the presentation

Previous data frame – Approach (Data cleaning)

- Data frame is cleaned following important steps
  - Total rows x columns (16,70,214x37)
  - Columns having missing values greater than 40% have been dropped
  - Analysis is performed on only 25 columns after dropping 12 columns
  - There are still 3 columns with missing values – AMT_Annuity, Amt_Goods_Price, CNT_payment,
- Data imputation
  - For AMT_Annuity, the missing values are replaced by “0” as most likely the annuity amount brought from previous loan application is “0”
  - For AMT_Annuity, Amt_Goods_Price missing values are not imputed by kept as missing values because any other way of imputation will imbalance the data and make analysis inaccurate.
- Data standardization
  - The numeric variables values were standardized to round off to 1 decimal places.

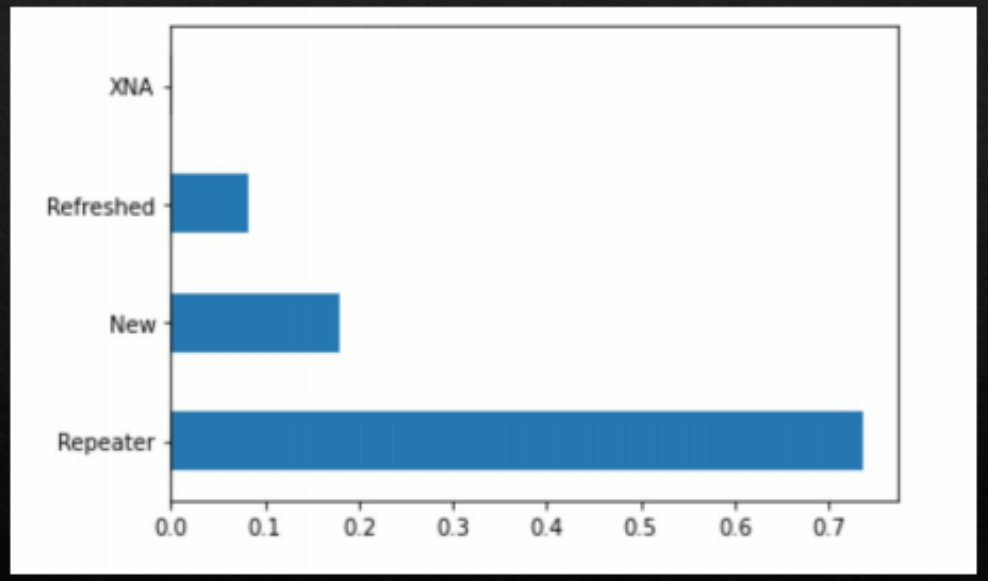
Univariate analysis (Categorical)

- Univariate analysis – categorical features
  - NAME_CONTRACT_TYPES



INFERENCE: CASH AND CONSUMER LOANS CONSTITUTE 88% BETWEEN THE

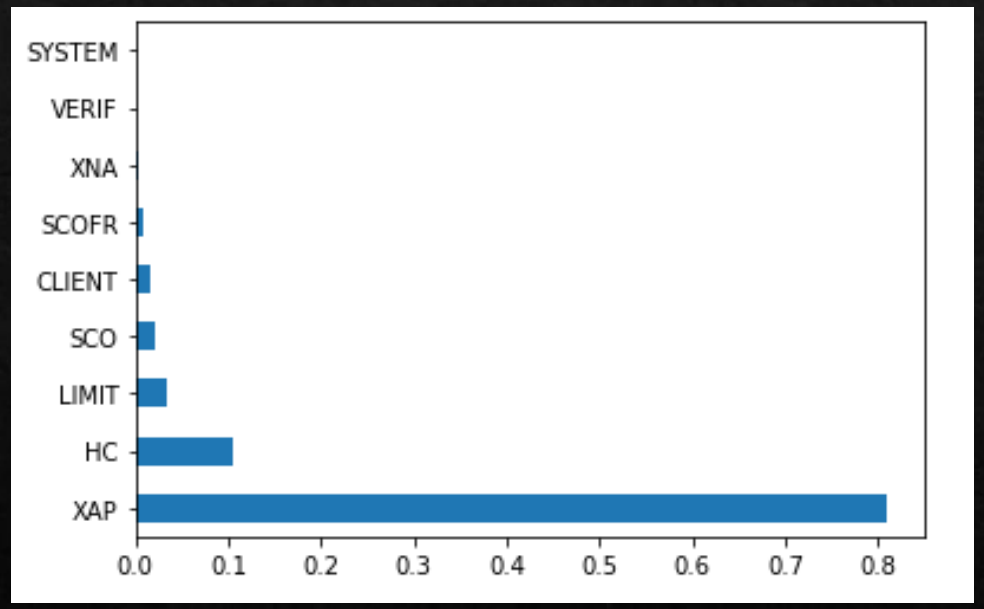
- Univariate analysis – categorical features
  - NAME_CLIENT_TYPE



INFERENCE: ABOUT 75% OF THE LOANS ARE REPEAT TYPE OF LOANS

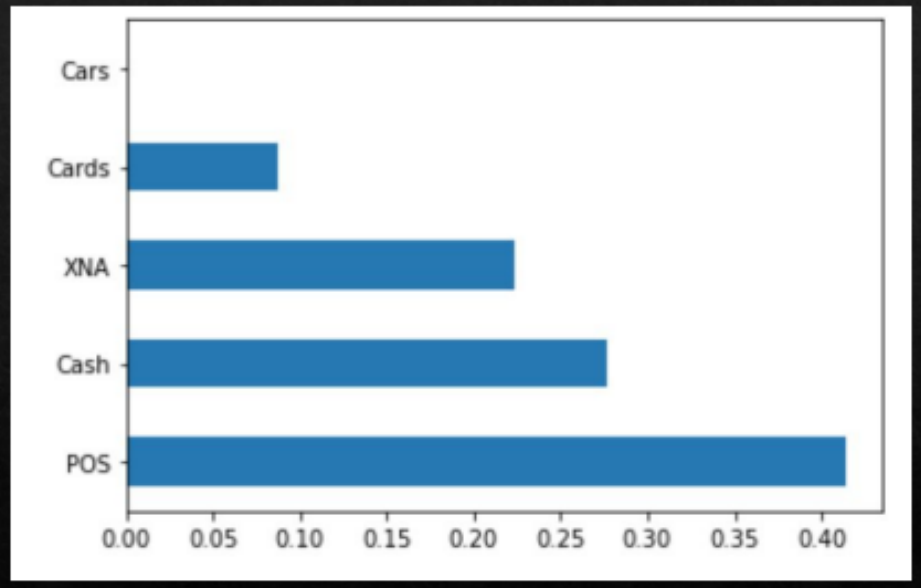
Univariate analysis (Categorical)

- Univariate analysis – categorical features
  - CODE_REJECT_REASONS



INFERENCE: XAP CONSTITUTES 80% OF THE REASONS WHY LOAN APPLICATIONS ARE GETTING REJECTED

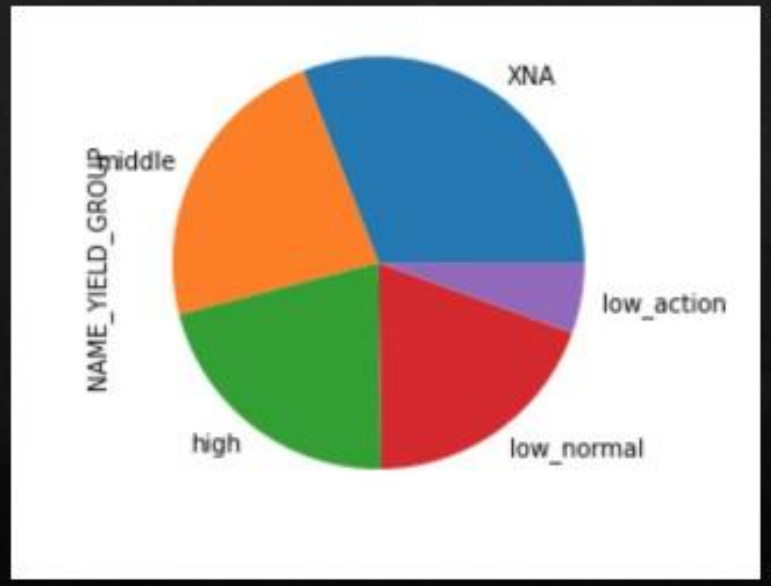
- Univariate analysis – categorical features
  - NAME_PORTFOLIO



INFERENCE: POS IS ABOUT 40% AND CASH IS ABOUT 28% WHICH ARE HIGHEST AOVE 5 CATEGORIES

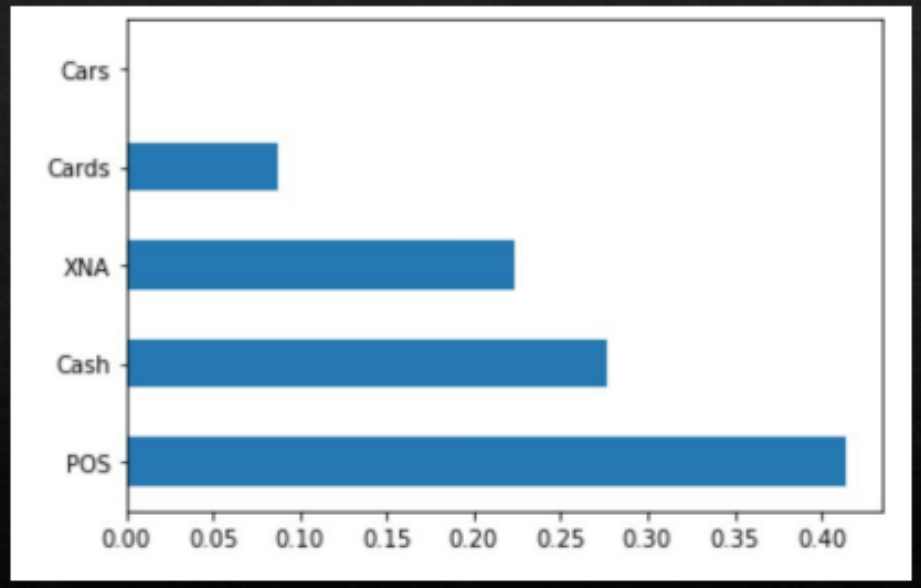
Univariate analysis (Categorical)

- Univariate analysis – categorical features
  - NAME_YIELD_GROUP



INFERENCE: GROUP INTEREST RATE ARE EQUALLY DISTRIBUTED ACROSS CATEGORIES. ONLY FOR LOW_ACTION ITS LOWER THAN OTHERS

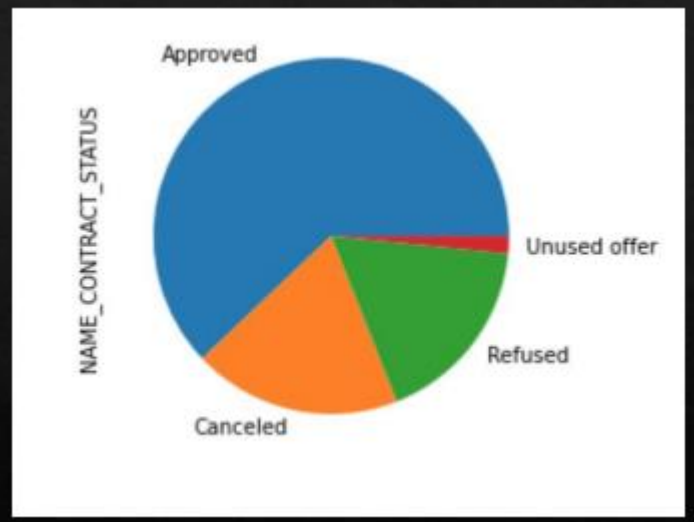
- Univariate analysis – categorical features
  - NAME_PORTFOLIO



INFERENCE: POS IS ABOUT 40% AND CASH IS ABOUT 28% WHICH ARE HIGHEST AOVE 5 CATEGORIES

Univariate analysis (Categorical)

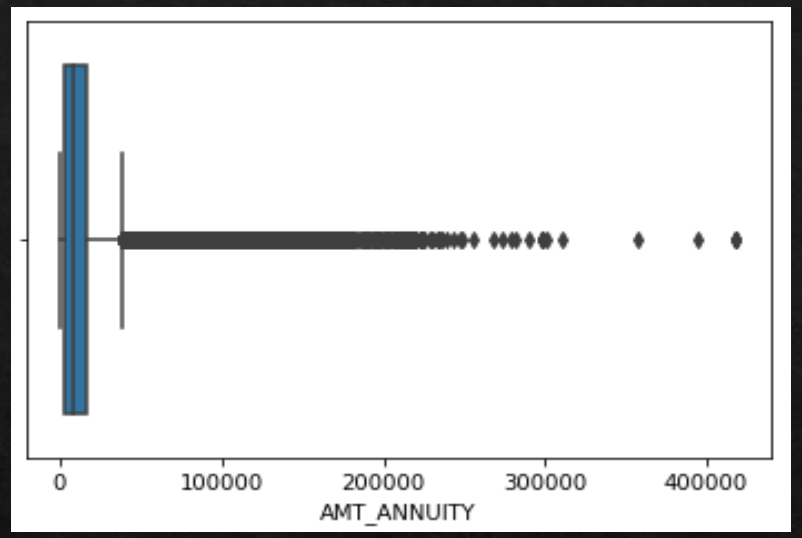
- Univariate analysis – categorical features
  - NAME_CONTRACT_STATUS

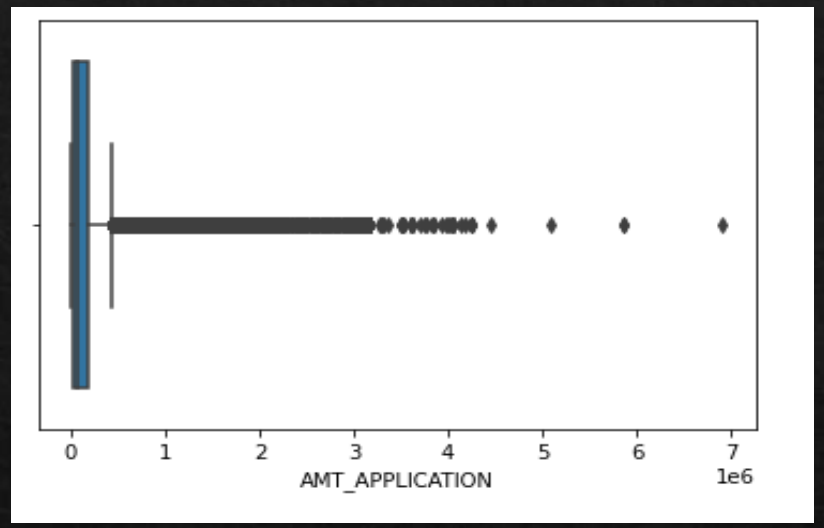


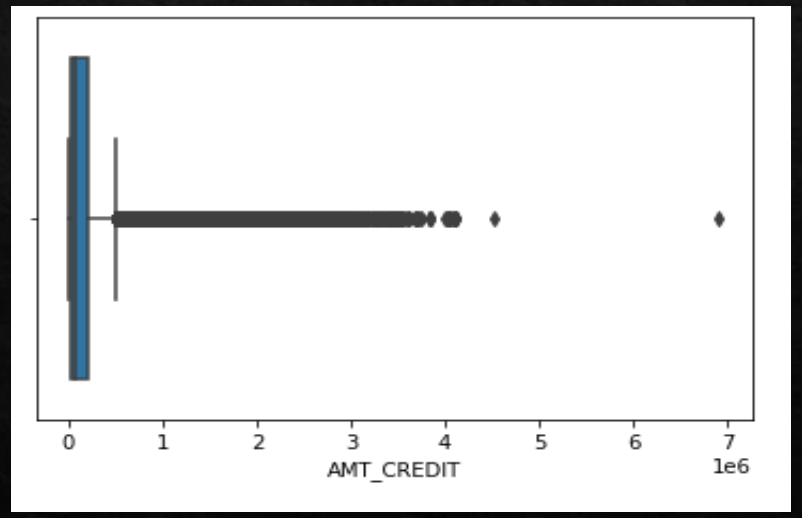
INFERENCE: APPROVED LOAN IS 62%, REFUSED LOAN IS 18.9% FOLLOWED BY CANCELED LOAN IS 17.4%. UNUSED OFFER IS AROUND 1%

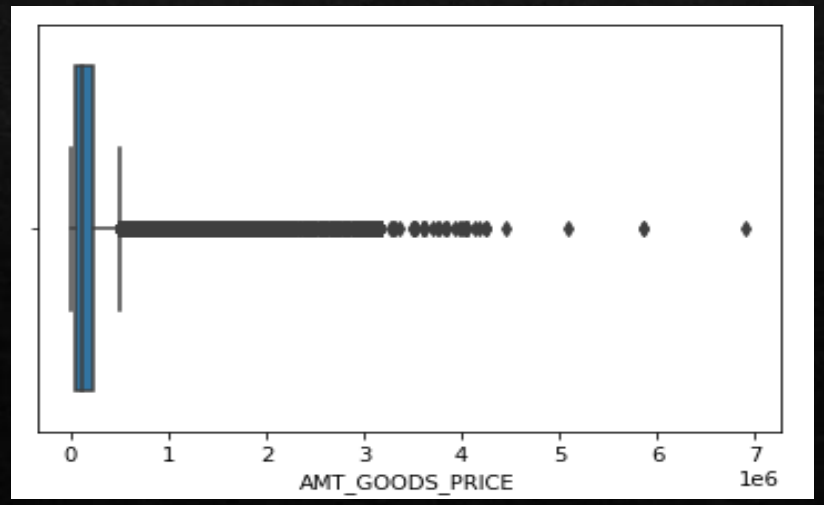
- Univariate analysis – categorical features
  - NAME_PORTFOLIO

Univariate (Numeric)









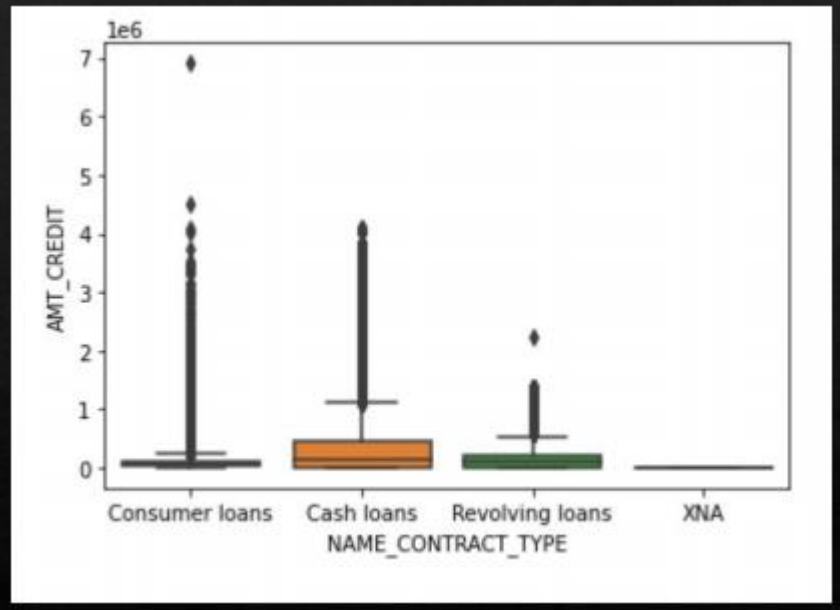
INFERENCE: WHEN WE ARE WORKING WITH ABOVE AMT RELATED NUMERIC FEATURES, WE WILL NOT BE WORKING WITH MEAN BUT QUANTILE OR PERCENTILE (NOT REMOVING OUTLIERS SINCE THEY CAN BE GENUINE AND REQUIRED FOR ANALYSIS OF OTHER VARIABLES)

Bivariate analysis (Numeric)

Numeric- Numeric
ALL AMOUNT VARIABLES

INFERENCE: THE BIVARIATE ANALYSIS IS PLOTTED FOR ALL NUMERIC VALUES AND THERE IS POSITIVE CORRELATION BETWEEN THE VARIABLES. CORRELATION MATRIX IS PLOTTED WHICH RESEALS THAT ITS HIGH BETWEEN 0.8 TO 1

Numeric- Categoric
CONTRACT_TYPE with AMT_CREDIT



INFERENCE: BOX PLOT IS NOT GIVING GREAT INSIGHT. HOWEVER AMOUNT CREDIT IS SPREAD OUT FOR CASH LOANS THAN CONSUMER AND REVOLVING LOANS.

CREATING TWO DATAFRAMES – FOR APPROVED AND REFUSED CATEGORIES

Numeric- Numeric
ALL AMOUNT VARIABLES – JUST FOR APPROVED CATEGORY

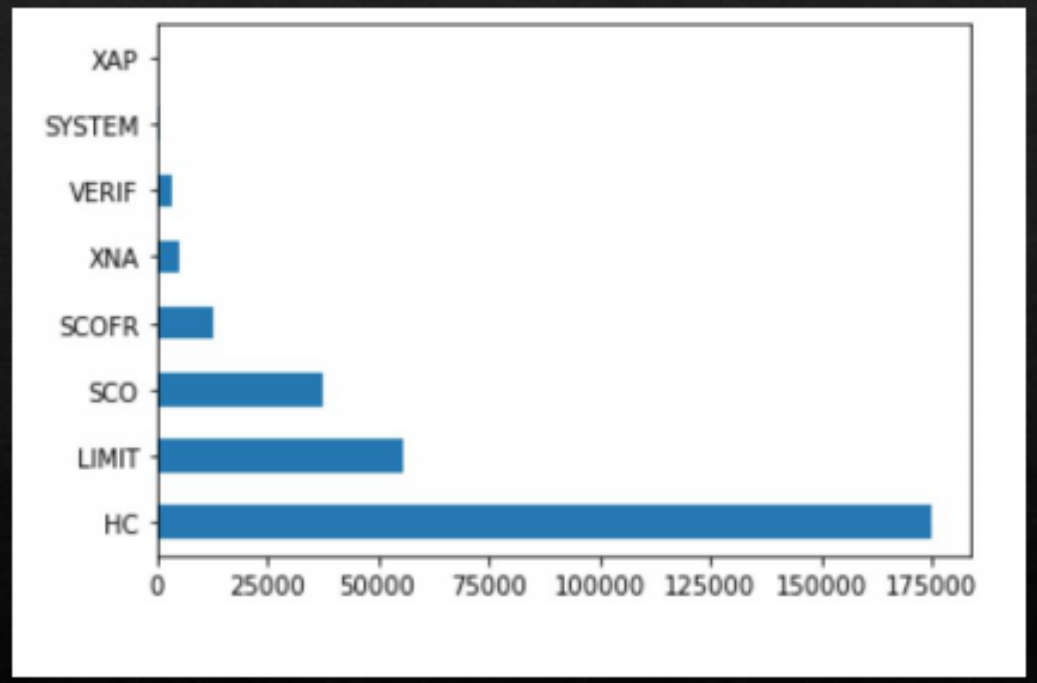
INFERENCE: CORRELATION MATRIX IS PERFORMED WHICH SHOWS HIGH CORRELATION FROM 0.81 TO 1

Numeric- Numeric
ALLAMOUNT VARIALES – JUST FOR REFUSED CATEGORY

INFERENCE: CORRELATION MATRIX IS PERFORMED WHICH SHOWS HIGH CORRELATION FROM 0.82 TO 1

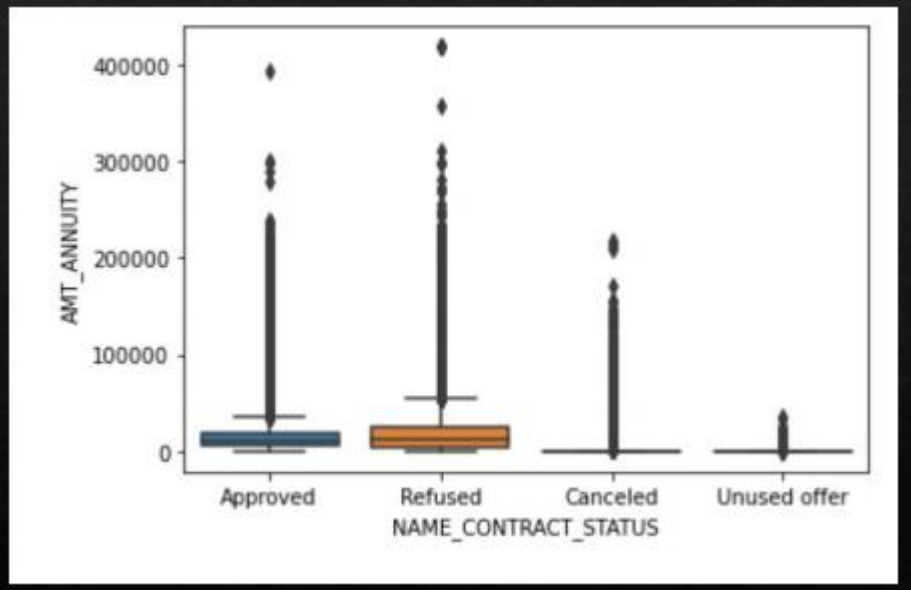
CREATING TWO DATAFRAMES – FOR APPROVED AND REFUSED CATEGORIES

CODE_REJECT_REASON
ANALYSIS OF REASONS BEHIND REJECTION OF LOAN APPLICATIONS



INFERENCE: 90% OF LOANS ARE GETTING REJECT BETWEEN THREE CATEGORES OF HC, LIMIT AND SCO

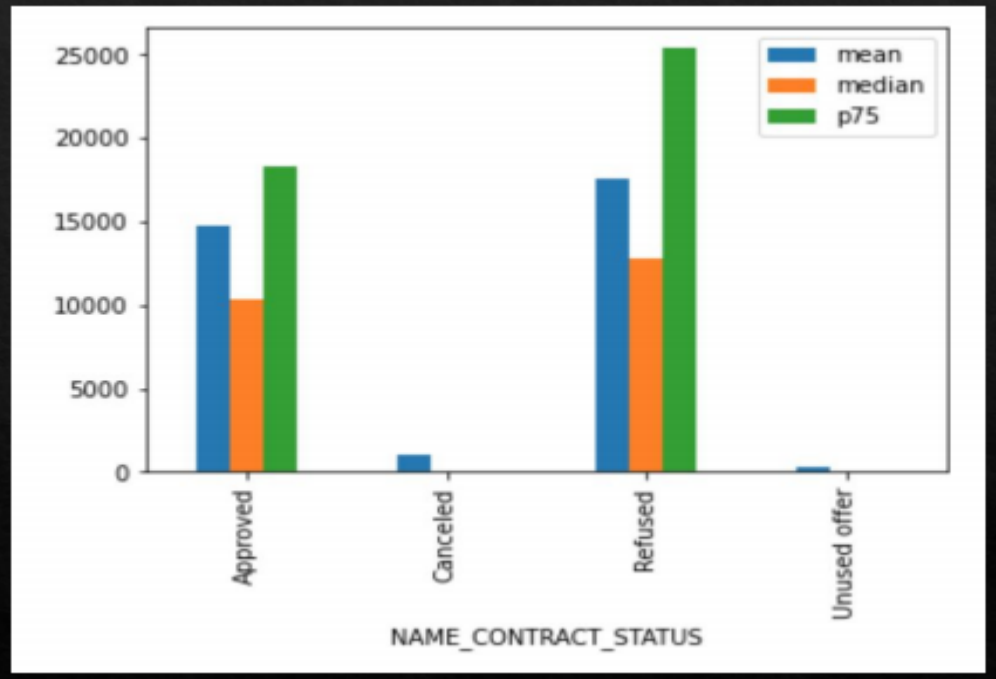
Numeric- Categoric
AMT_ANNUITY WITH NAME_CONTRACT_ STATUS



INFERENCE: CANCELLED AND UNUSED OFFERS HAVE TYPICALLY VERY LOW AMT_ANNUITY

CREATING TWO DATAFRAMES – FOR APPROVED AND REFUSED CATEGORIES

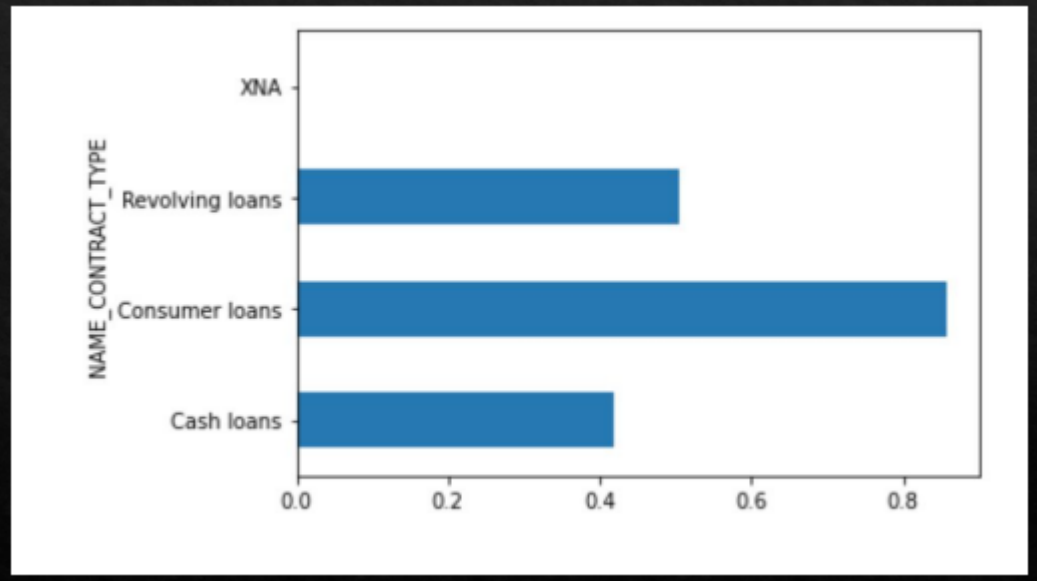
ANALYSIS FOR 75TH PERCENTILE AS OUTLIERS ARE CONTINOUS BUT VERY HIGH VALUES



INFERENCE:AMT_ANNUITY IS TYPICALLY HIGHER IN CASE OF LOANS WHICH ARE GETTING REFUSED THAN THAT OF LOANS THAT WHICH ARE GETTING ACCEPTED

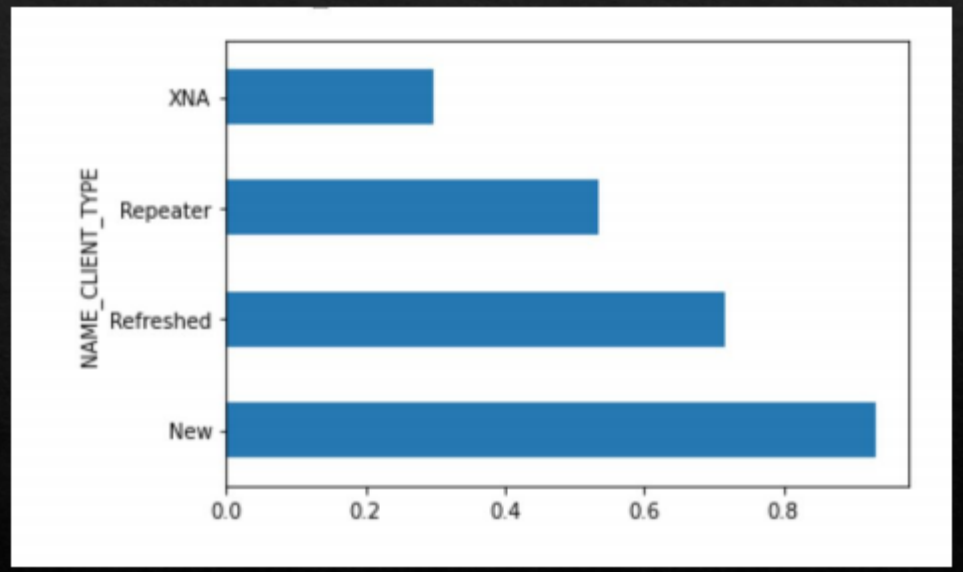
ANALYSIS OF TARGET VARIABLE BY DEFINING APRPOVAL RATE (FROM NAME CONTRACT STATUS) W.R.T. TO ALL OTHER IMP VARIABLES

APPROVAL RATE WITH NAME_ CONTRACT_ TYPE



INFERENCE:THE APPROVAL RATE OF CONSUMER LOANS IS HIGHEST FOLLOWED BY REVOLVING LOAN AND THEN BY CASH LOANS

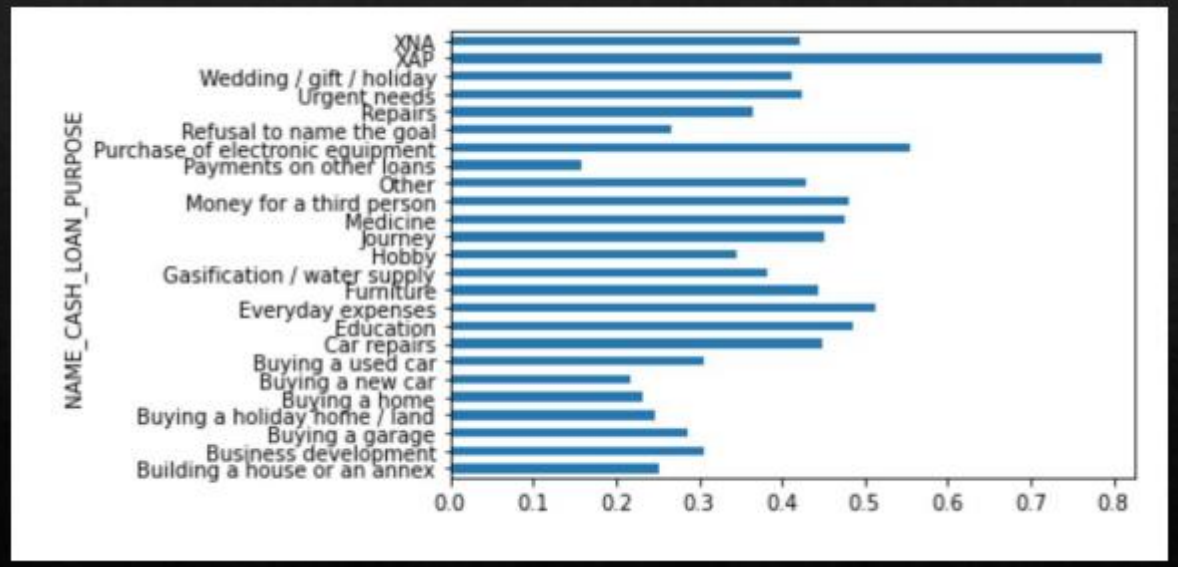
APPROVAL RATE WITH NAME_CLIENT_TYPE



INFERENCE:NEW CLIENTS' LOAN APPLICATION ARE LIKELY TO BE APPROVED HIGHEST FOLLOWED BY REFRESHED AND THEN REPEATER

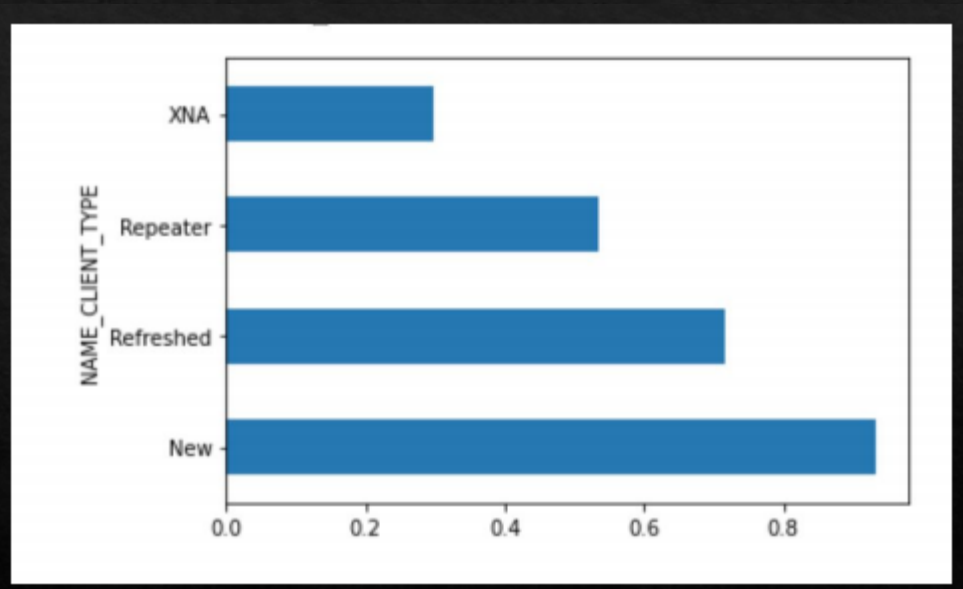
ANALYSIS OF TARGET VARIABLE BY DEFINING APRPOVAL RATE (FROM NAME CONTRACT STATUS) W.R.T. TO ALL OTHER IMP VARIABLES

APPROVAL RATE WITH NAME_ CASH_LOAN_PURPOSE



INFERENCE:: XAP LOANS HAS GOT THE HIGHER APPROVAL RATE FOLLOWED BY PURCHASE OF ELECTRONIC EQUIPMENTS. REST ARE AROUND SAME APPROVAL RATE

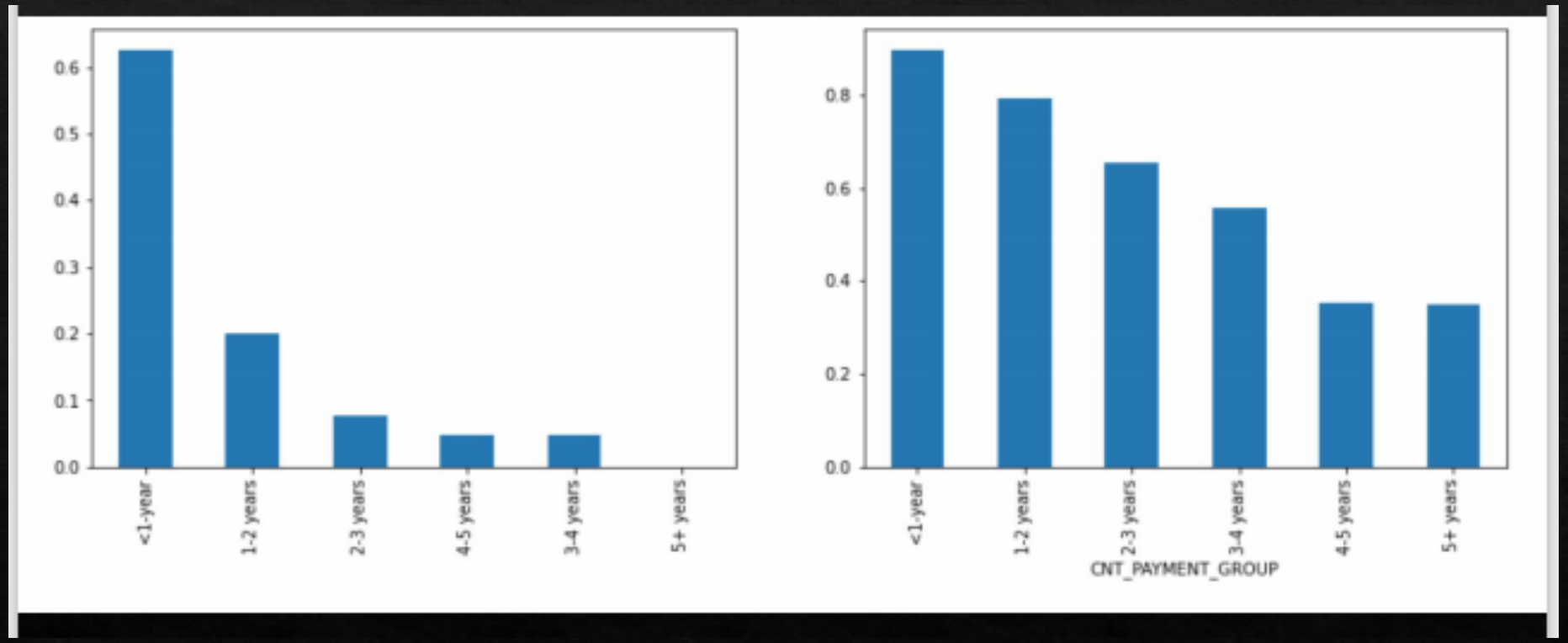
APPROVAL RATE WITH NAME_CLIENT_TYPE



INFERENCE:NEW CLIENTS' LOAN APPLICATION ARE LIKELY TO BE APPROVED HIGHEST FOLLOWED BY REFRESHED AND THEN REPEATER

ANALYSIS OF TARGET VARIABLE BY DEFINING APRPOVAL RATE (FROM NAME CONTRACT STATUS) W.R.T. TO ALL OTHER IMP VARIABLES

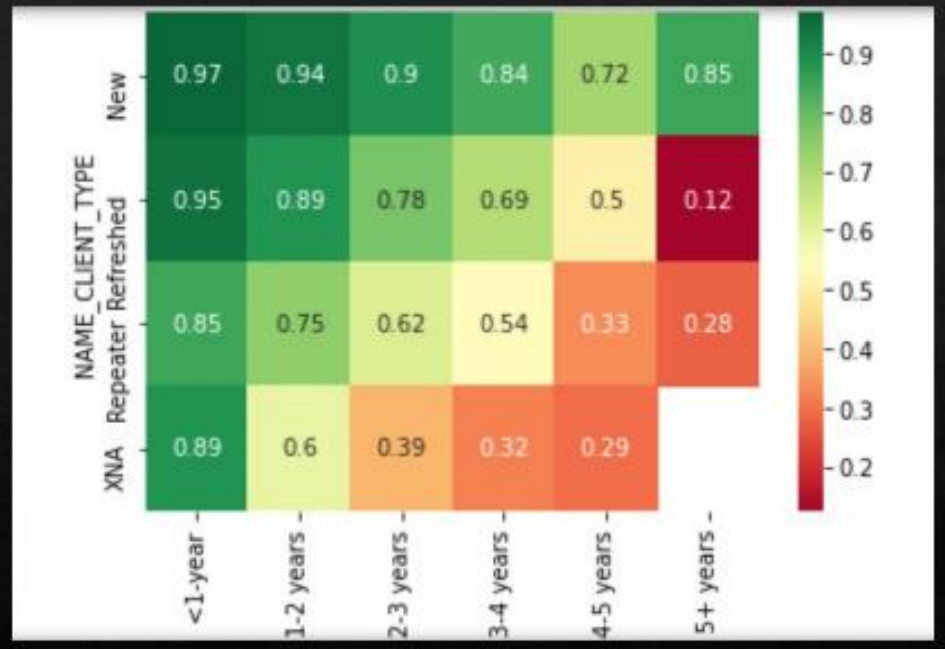
APPROVAL RATE WITH NAME_ CNT_PAYMENT GROUP
TERM OF CONTRACT OR CNT_PAYMENT HAS BEEN BINNED INTO 6 BUCKETS



INFERENCE:AS THE TERM OF LOAN INCREASES THE APPROVAL RATE DECRESES. FOR LOAN WITH TERM UPTO 1 YEAR, THE APPROVAL RATE IS > 80% WHICH IS HIGHEST AMONG ALL THE CNT_PAYMENT (TERM) CATEGORIES

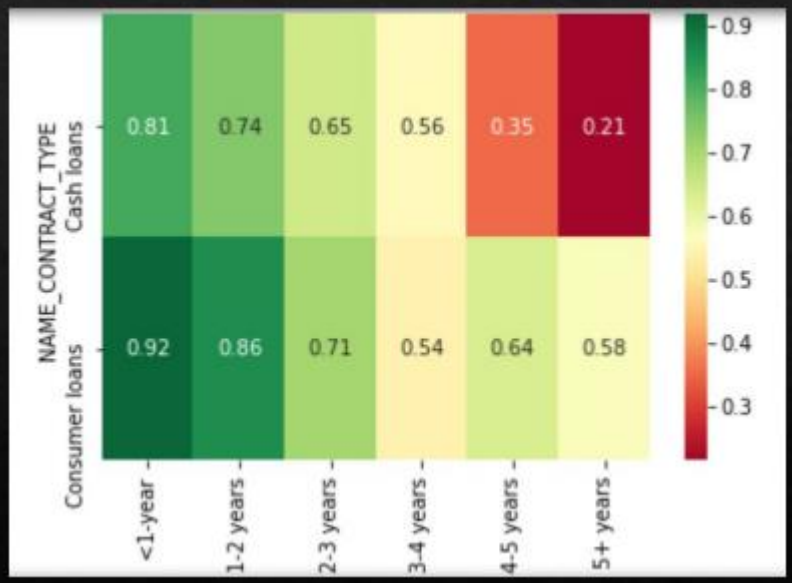
MULTIVARIATE ANALYSIS OF TARGET VARIABLE W.R.T. TO ALL OTHER IMP VARIABLES

APPROVAL RATE WITH NAME_ CLIENT_TYPE AND CNT_PAYMENT_GROUP (TERM)



INFERENCE:: THE APPROVAL RATE FOR ANY CLIENT TYPE OF LOAN DECREASES WITH HIGHER TERM (CNT_PAYMENT) OF LOANS. IN OTHER WORDS THE SHORTER THE TERM OF LOAN, HIGHER IS THE APPROVAL RATE.

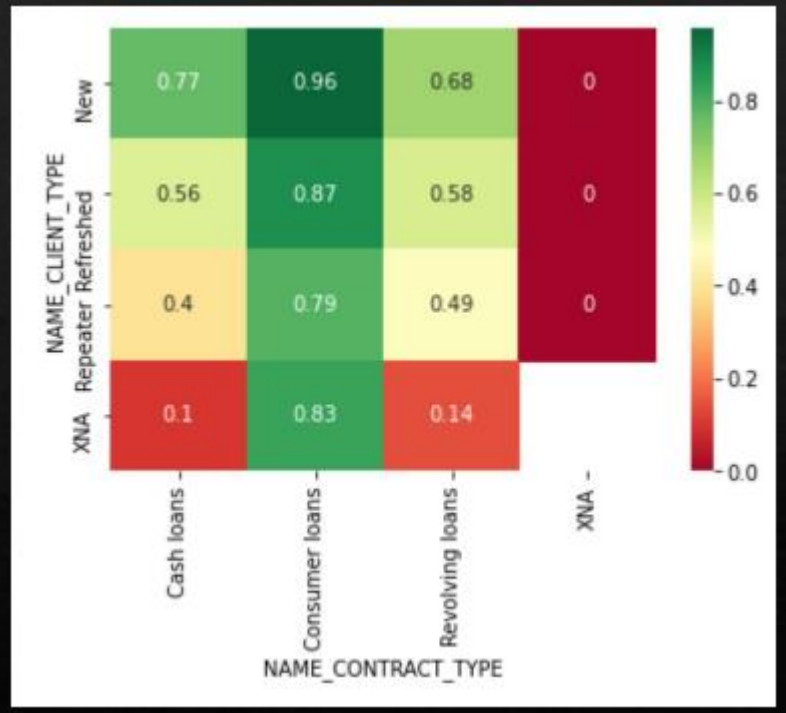
APPROVAL RATE WITH NAME_CONTRACT_TYPE AND CNT_PAYMENT_GROUP



INFERENCE:APPROVAL RATE DECREASES FOR ANY CONTRACT_TYPE AS THE TERM (CNT_PAYMENT) INCREASES. ALSO AS THE TERM INCREASES CASH LOANS HAVE LESS APPROVAL RATE THAN CONSUMER LOANS.

MULTIVARIATE ANALYSIS OF TARGET VARIABLE W.R.T. TO ALL OTHER IMP VARIABLES

APPROVAL RATE WITH NAME_ CLIENT_TYPE AND NAME_CONTRACT_TYPE



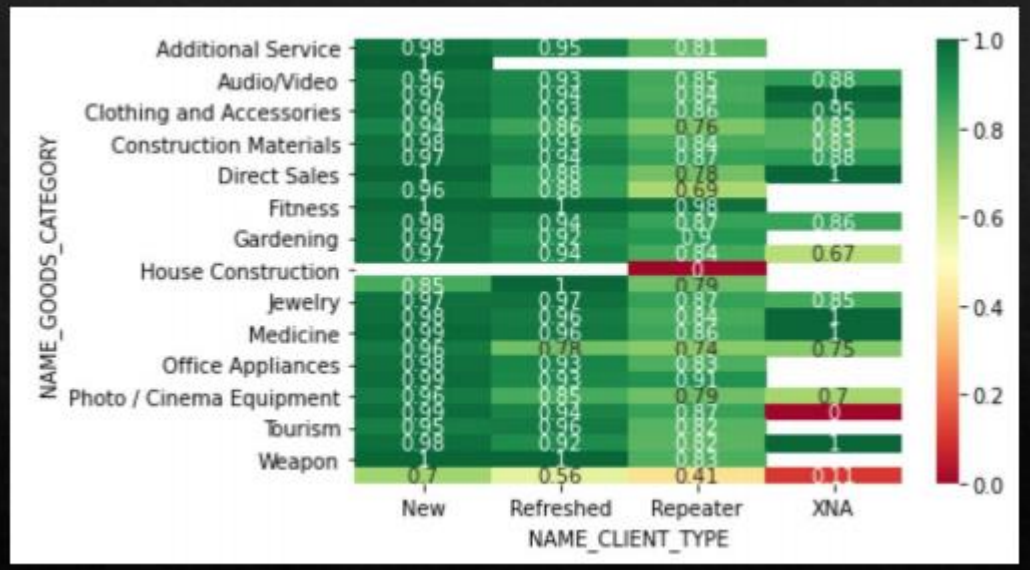
INFERENCE: APPLICANTS WHO ARE APPLYING FOR LOAN FOR FIRST TIME (NEW) IN CONSUMER LOAN (CONTRACT_TYPE) HAS THE HIGHEST APPROVAL RATE. INFERENCE: SIMILALRY REVOLVING LOANS TYPE IN XNA CATEGORY HAS THE LOWEST APPROVAL RATE

APPROVAL RATE WITH NAME_CASH_LOAN_PURPOSE AND NAME_CLIENT_TYPE

INFERENCE:WITHIN CASH TYPE OF LOANS, NEW HAS HIGHEST APPROVAL RATE AND IT DECREASES PROGRESSIVELY FOR REFRESHED, REPEATER TYPE OF LOANS AND SO ON. INFERENCE: THERE ARE CERTAIN CATEGORIES LIKE HOBBY, MEDICINE AND MONEY FOR THIRD PERSON HAS 100% APPROVAL RATE IN DIFFERNT CLIENT TYPES

MULTIVARIATE ANALYSIS OF TARGET VARIABLE W.R.T. TO ALL OTHER IMP VARIABLES

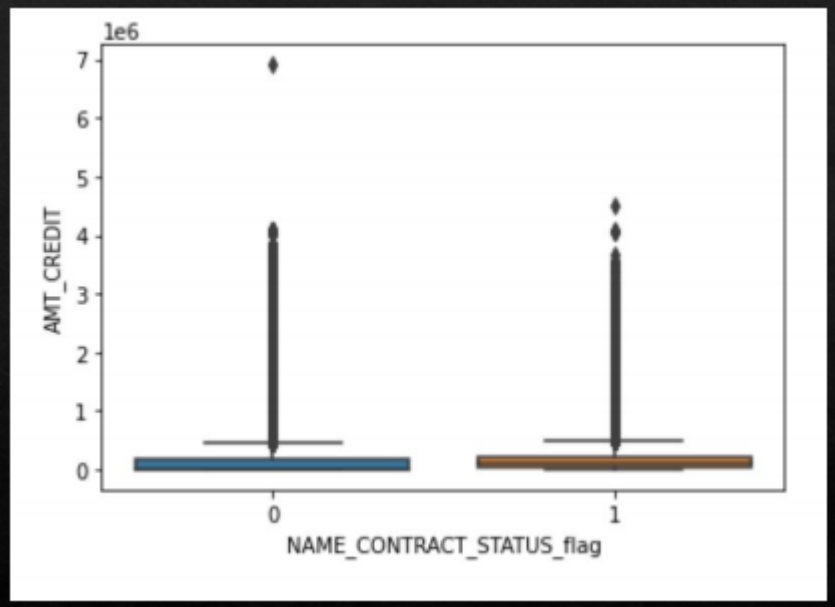
APPROVAL RATE WITH NAME_ GOODS_CATEGORY AND NAME_CLEINT_TYPE



INFERENCE: FOR ALL TYPES OF GOODS CATEGORIES, THE APPROVAL RATE IS HIGH FOR NEW CLIENT AND PROGRESSIVELY DECREASES WITH REFRESHED AND REPEATER. INFERENCE: FOR HOUSE CONSTRUCTION FOR REPEATER TYPE OF LOANS THE APPROVAL RATE IS LOWEST OR "0". FOR FITNESS AND WEAPON CATEGORIES THE APPROVAL RATE ARE VERY HIGH (2 OUT OF 3 CATEGORIES HAVE 100%

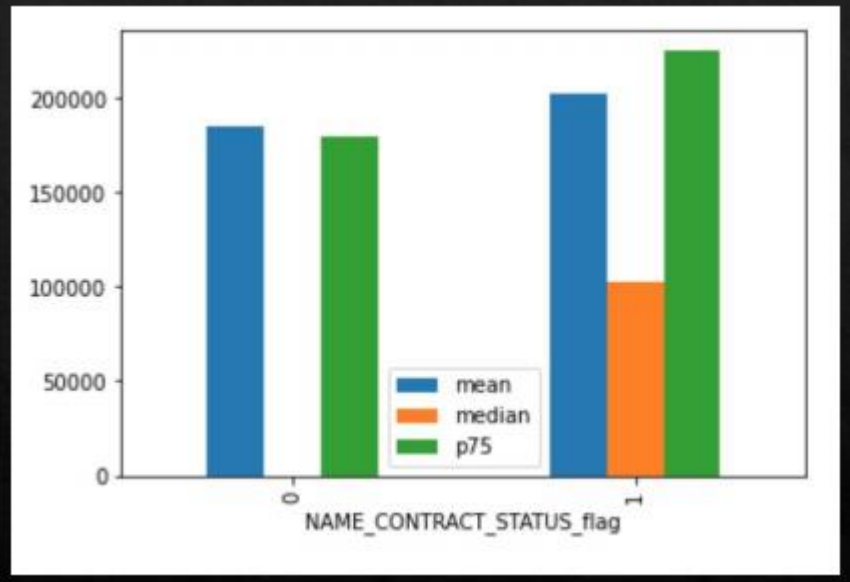
ANALYSIS OF TARGET VARIABLE W.R.T. AMT_CREDIT

APPROVAL RATE WITH AMT_CREDIT



INFERENCE: DUE TO HIGH VALUE ITS DIFFICULT TO INTERPRET FROM THE BOX PLOT. HENCE, WE WILL LOOK AT QUANTILES (75%) ALSO. DIRECTION: DEFINING FUNCTION FOR 75TH QUANTILE.

SINCE BOX PLOT IS NOT EASY TO COMPREHEND, WE ARE TAKING MEDIAN AND 75TH QUANTILE FOR ANALYSIS



INFERENCE: FOR LOAN THAT WERE NOT APPROVED"0"(REFUSED, CANCELED, UNUSED OFFERS) - ITS SEEMS HIGH AMT_CREDIT GOT REFUSED WHEREAS MEDIUM (MEDIAN) TO LOW AMOUNT (MAY INCLUDE 0 ALSO) ARE EITHER CANCELLED OR UNUSED OFFERS.
INFERENCE: FOR LOAN THAT WERE APPROVED"1" - THE AMT_CREDIT ARE WELL SPREAD OUT WITH VALUES HIGH ; MEDIUM ARE LOW ARE GETTING APPROVED

Application data – Data cleaning approach

Cleaning the Missing Data and Deriving Columns

- Removed 58 columns having more then 30% null values, as these columns can skewed our analysis.
- We remain with 64 columns in dataframe for further analysis.
- Removing unwanted 27 columns from the data frame.
- Now lets looks at columns having less then 30% null values -
  - Goods price amount have 278 records as null values and as volume is very less 0.1%, we can impute with mean of this column.
  - There are two records have null number of family members and majority records are having 2 family members, so lets impute null values with 2 to take care null records in this column.
  - Annuity amount have 12records as null values and as volume is very less 0.1%, we can impute with mean of this column
- Derived
  - Derived “age” and “year employed” column from DAYS_BIRTH and DAYS_EMPLOYED columns.
  - Binning for continuous variable columns (AMT_INCOME_TOTAL and AMT_CREDIT).

Data Insights, Quality and remove Outliers

Features	Findings
TARGET	• 92% of data is belongs to non-difficulties
NAME_CONTRACT_TYPE	• NAME_CONTRACT_TYPE column looks good and distributed among Cash and Revolving loans. • 90% of data is belongs to Cash loans
CODE_GENDER	• Female applicant are twice with respect to male applicant. • Imputed “XNA” 'CODE_GENDER' with "F" as we have clear majority for Female applicants (~65%).
FLAG_OWN_CAR	• This feature says 65% of applicant own the car.
FLAG_OWN_REALTY	• This feature says ~70% of applicant own real state property.
YEARS_EMPLOYED	• Removed outliers greater than -100 years.
AMT_ANNUITY	• We dropped the records above 60,000 to remove unusually high amount annuity price from the dataframe.
AMT_CREDIT	• We dropped the records above 1,500,000 to remove unusually high amount credit price from the dataframe.

Continue…

Feature	Findings
CNT_CHILDREN	• After observing outliers, we can conclude to remove CNT_CHILDREN more then 2 as 70%,19% and 8% applicant population belongs to 0,1,2 respectively.
CNT_FAM_MEMBERS	• 52.22% values belongs to 2 family members.
NAME_INCOME_TYPE	• 52% applicant population belongs to Working income class type.
NAME_EDUCATION_TYPE	• 71% applicant population belongs to Secondary / secondary special education type.
NAME_FAMILY_STATUS	• ~64% applicant population belongs to Married class and this is highest. • Imputed unknown to married as 64% applicant population belongs to Married family status.
NAME_HOUSING_TYPE	88% applicant population belongs to House / apartment housing type.
AMT_INCOME_TOTAL	• We dropped the records above 1,000,000 to remove unusually high amount credit price from the dataframe.
AMT_INCOME_TOTAL	• We dropped the records above 300,000 to remove unusually high amount income from the dataframe.

Continue…

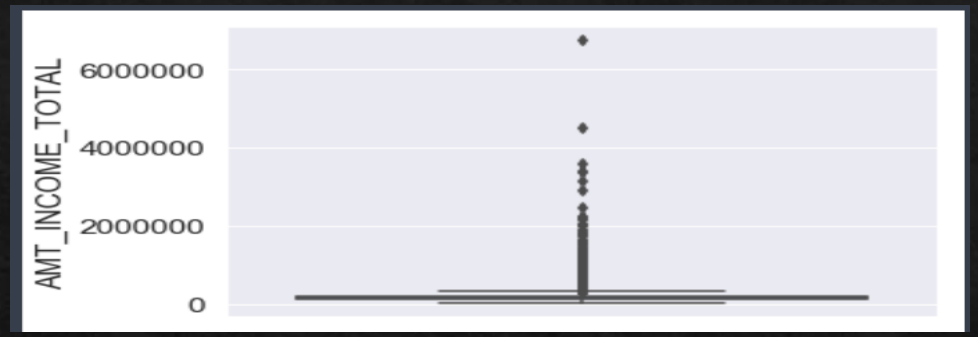
Feature	Findings
ORGANIZATION_TYPE	• Imputed XNA to Other values
AGE	• AGE data looks consistence and spread between 35 to 55 age for loan applicants.
AMT_GOODS_PRICE	• We dropped the records above 1,400,000 to remove unusually high amount goods price from the dataframe.

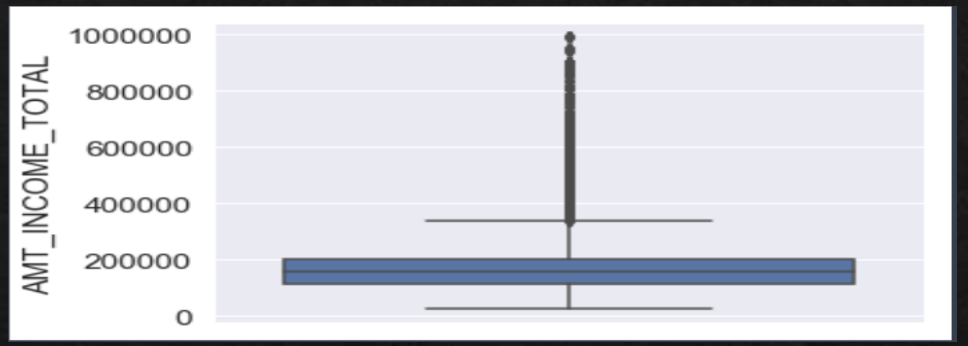
Observed few binary columns, those are divided into either 1 or 0, following is the column list –

- REG_REGION_NOT_LIVE_REGION
- REG_REGION_NOT_WORK_REGION
- LIVE_REGION_NOT_WORK_REGION
- REG_CITY_NOT_LIVE_CITY
- REG_CITY_NOT_WORK_CITY
- LIVE_CITY_NOT_WORK_CITY

Outlier example

AMT_INCOME_TOTAL





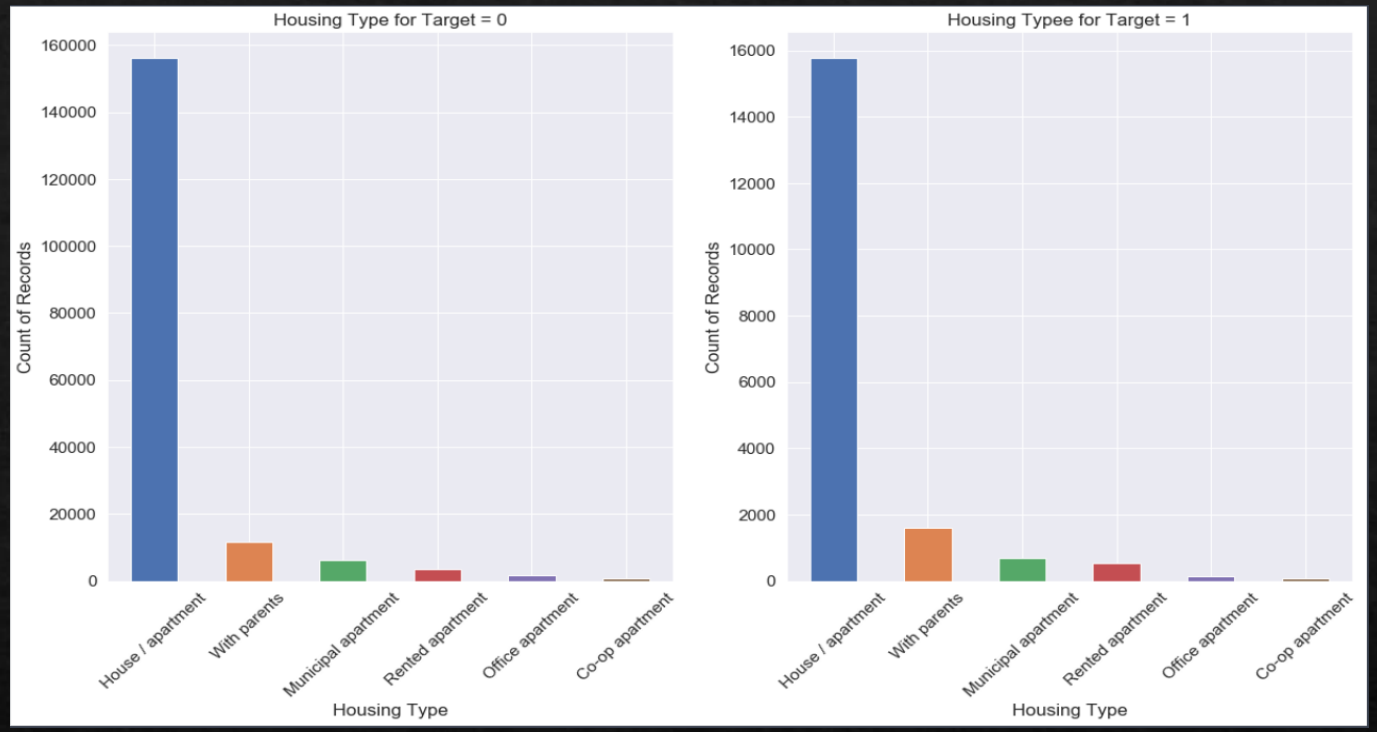

```
print(df_application_data.shape)
df_application_data = df_application_data[~(df_application_data['AMT_INCOME_TOTAL']>300000)]
print(df_application_data.shape)

sns.boxplot(x=df_application_data['AMT_INCOME_TOTAL'],orient='v')
plt.show()
```

```
(212723, 31)
(198618, 31)
```

Univariate Analysis

- . Categorical Variables –
  - CODE_GENDER
  - FLAG_OWN_REALTY
  - NAME_INCOME_TYPE
  - NAME_EDUCATION_TYPE
  - NAME_FAMILY_STATUS
  - NAME_HOUSING_TYPE
  - CNT_CHILDREN

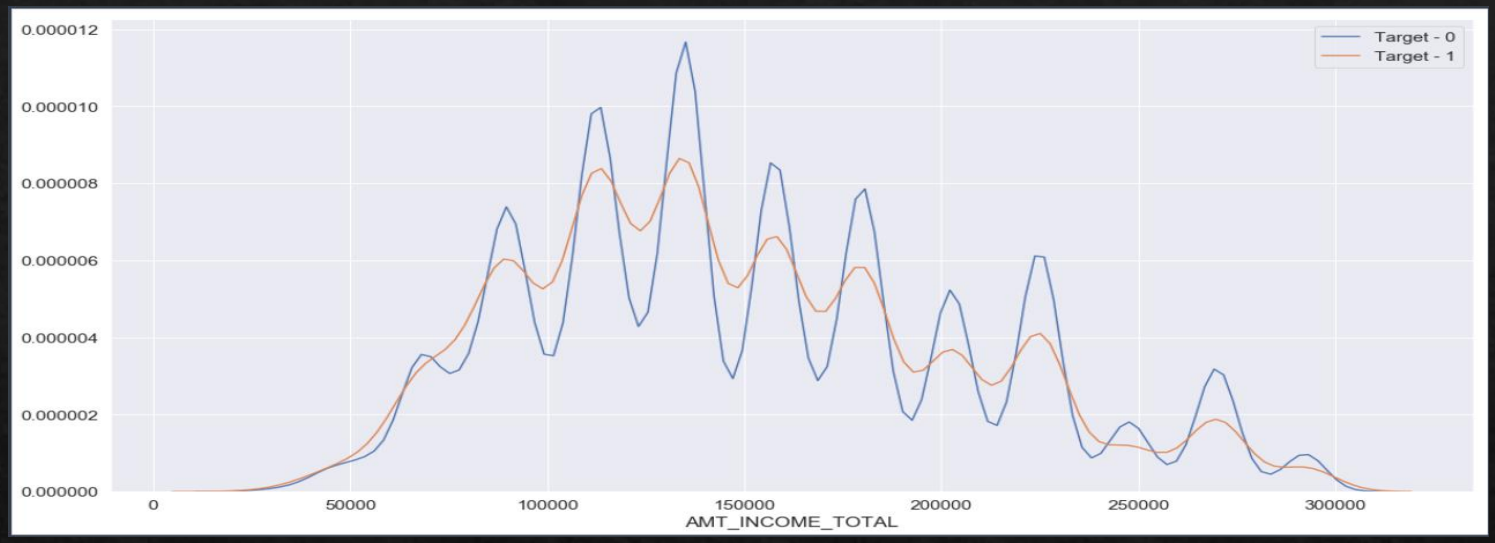
Right image is one of the example



- Inferences –
  - These features are almost having same ratio with respect to TARGET variable 0 and 1.

Univariate Analysis Continue..

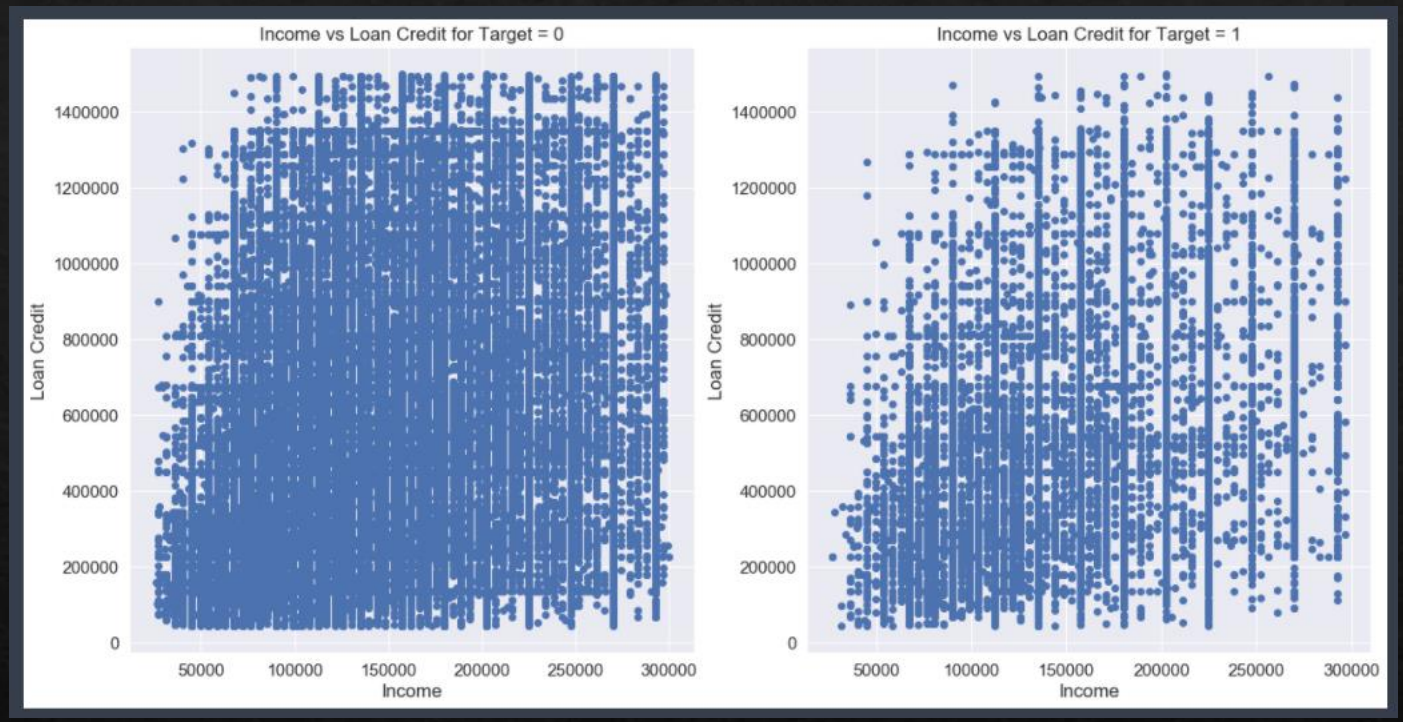
- Continuous Variables
  - AMT_INCOME_TOTAL
  - AMT_ANNUITY
  - AMT_GOODS_PRICE
  - YEARS_EMPLOYED
  - AGE



- Inferences –
  - Loan applicant with lower income are more likely to face payments difficulties.
  - Loan applicant with higher AMT_CREDIT will have difficulties for loan payments.
  - This looks similar but on slight note we can observe Loan applicant for higher AMT_GOOD_PRICE have difficulties for loan payments.
  - This looks similar but on slight note we can observe Loan applicant with higher YEARS_EMPLOYED have difficulties for loan payments.
  - This looks similar but on slight note we can observe Loan applicant with higher age have difficulties for loan payments.

Bivariate Analysis

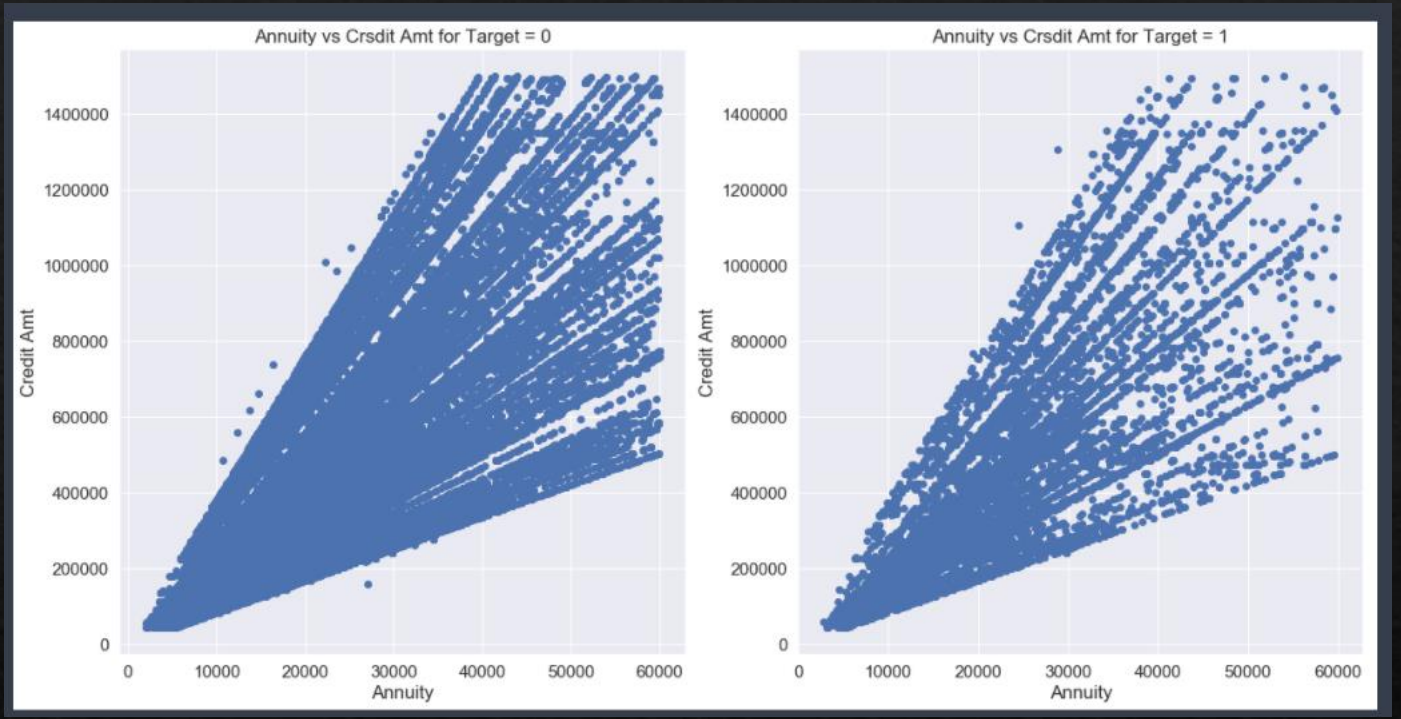
- Continuous - Continuous
  - AMT_INCOME_TOTAL vs AMT_CREDIT



- Inferences – Applicant with consistence income are most likely comfortable with loan payments.

Bivariate Analysis Continue..

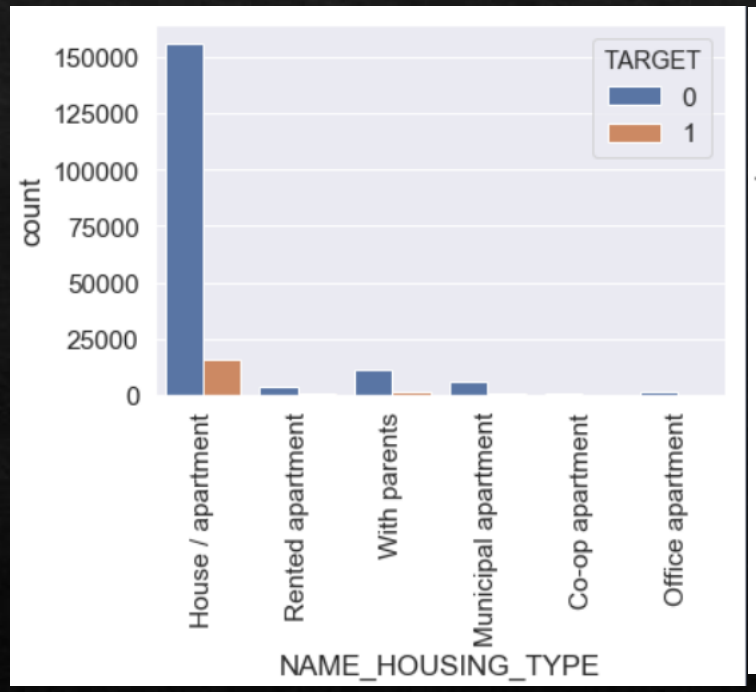
- Continuous - Continuous
  - 'AMT_CREDIT‘ and 'AMT_ANNUITY

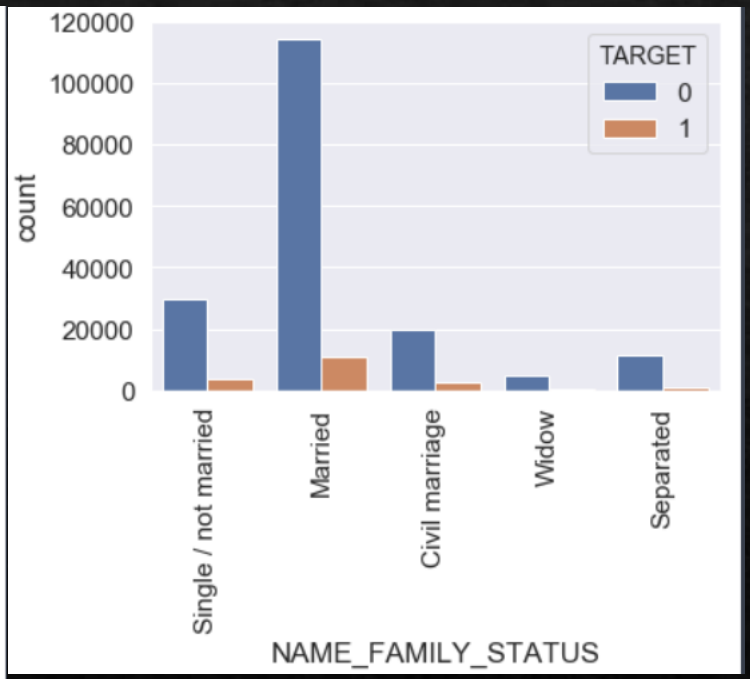


- Inferences – Amount credit and amount annuity are most likely in ratio but we can conclude applicant with higher annuity are most likely to get high amount credit and conformable with loan payments.

Bivariate Analysis Continue..

- Categorical –Categorical
  - NAME_INCOME_TYPE vs AMT_CREDIT





- Inferences – Married and living in house/apartment applicant are more comfortable for paying loan.


Thank
You