Exploratory Data Analysis for Bank Credit Dataset

By - Aditya Vijay Punjabi

Problem Statement

- There are two types of risk involved prior to disbursement of any loan request :
- A -: If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- B - : If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Objective

- This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

Steps Taken

- Data Importing
- Data Cleaning
- Detection of Outliers
- Check for data imbalance
- Univariate Analysis
- Bivariate Analysis
- Heatmap
- Insights From The Dataset

Data Importing

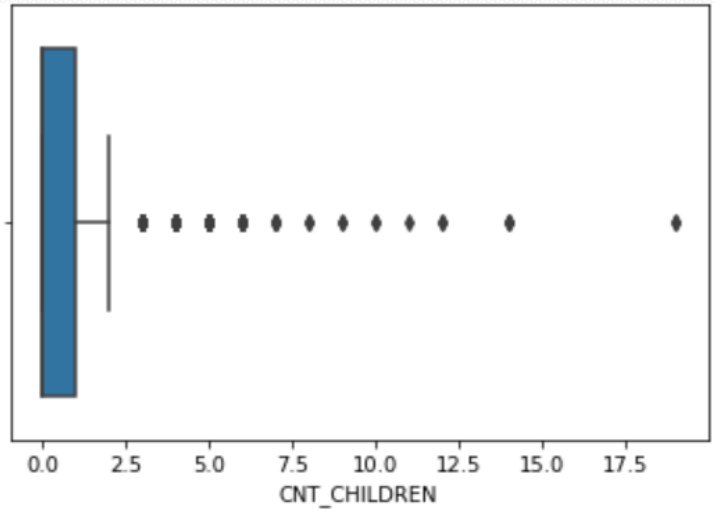
- There are 3 files provided
- 1. **application_data.csv** – In this file all the applications provideded by the applicants are given.
- 2. **previous_application.csv** – In this file the previous loan taken by the applicants are given.
- 3. **column_description.csv** – In this file the brief info about columns is given.

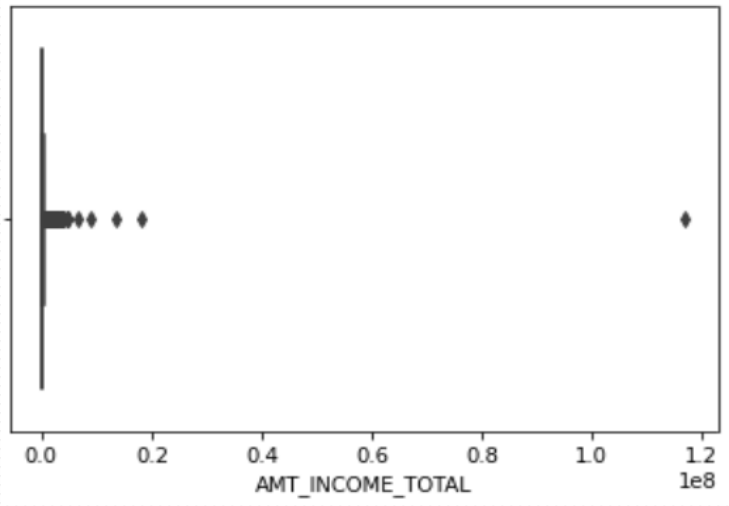
For Application_data.csv
Data Cleaning

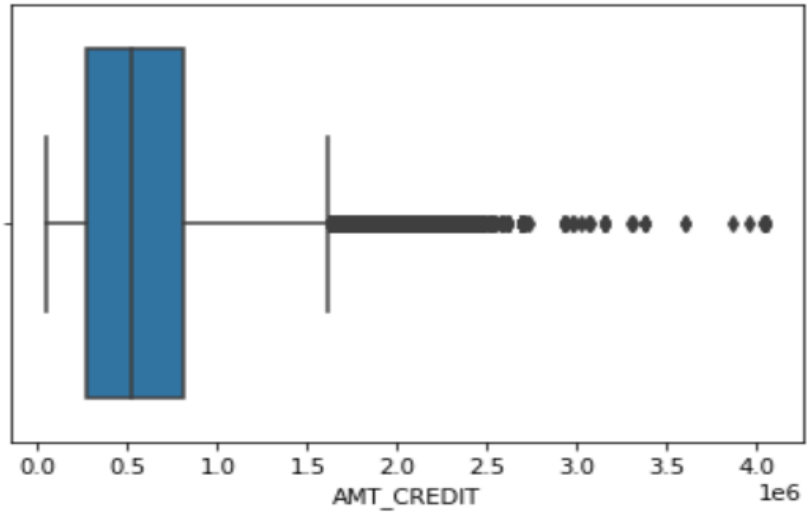
- Firstly checked the shape and info about the dataset.
- After that dropped the columns which are having missing data more than 45%.
- After that imputed the missing values for the columns.
- Then corrected the data types of columns.

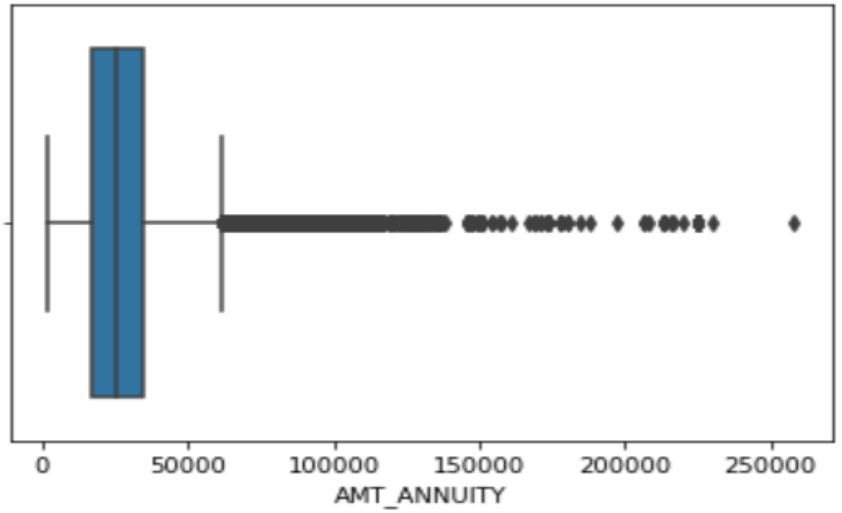
Detection of Outliers

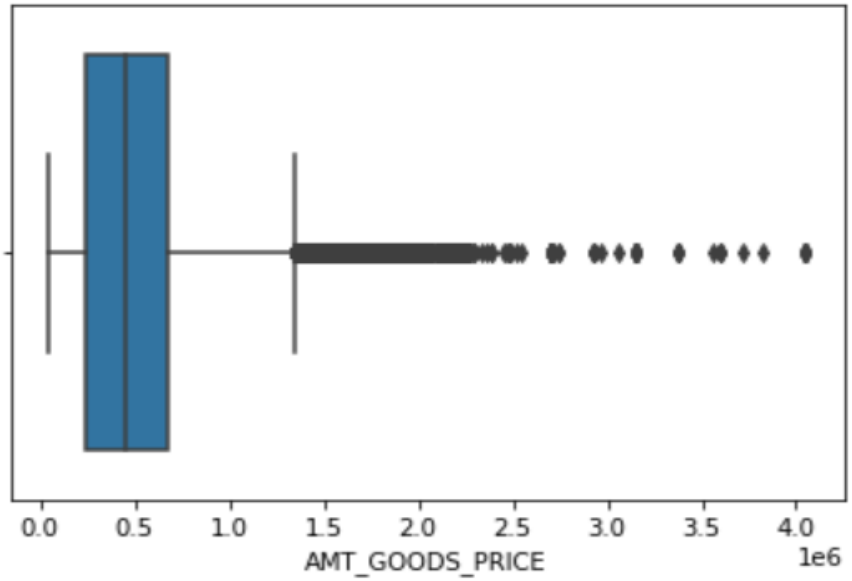
- In the numerical columns the most of the columns are having outliers one or two columns are having no outliers.
- Columns having outliers.

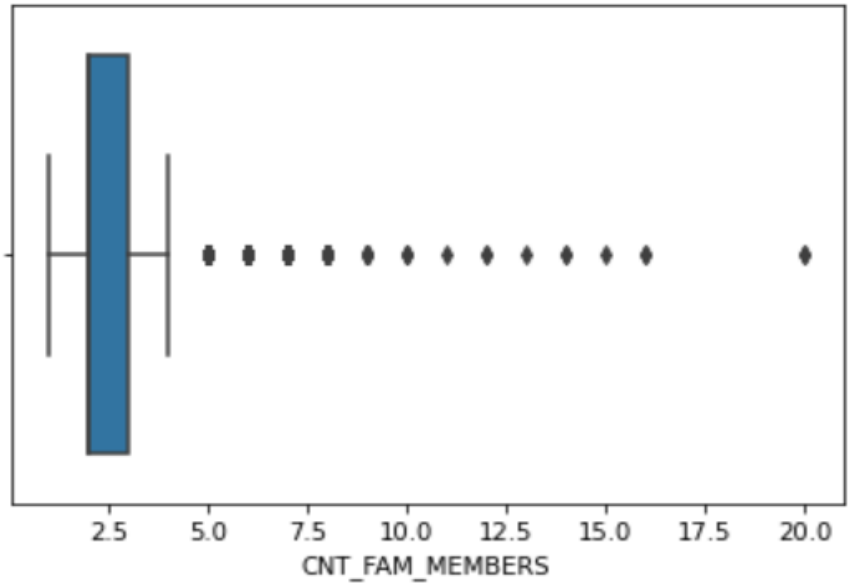




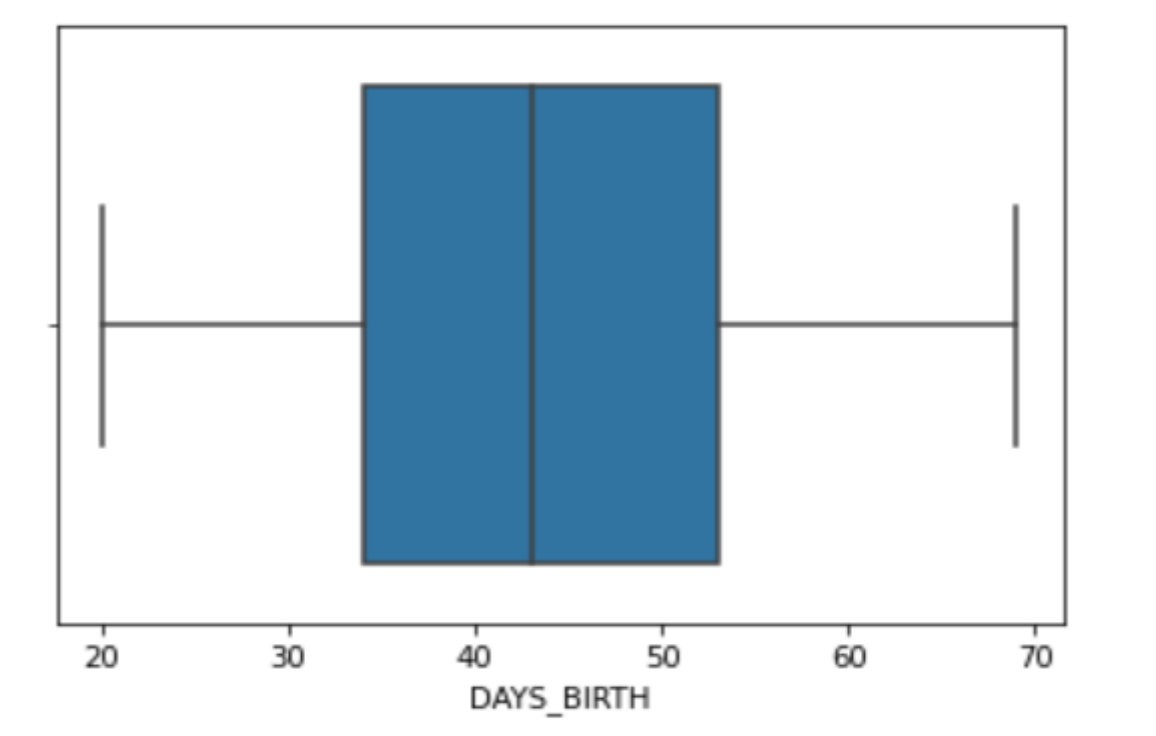

0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
AMT_CREDIT
1e6


0
50000
100000
150000
200000
250000
AMT_ANNUITY


0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
AMT_GOODS_PRICE
1e6

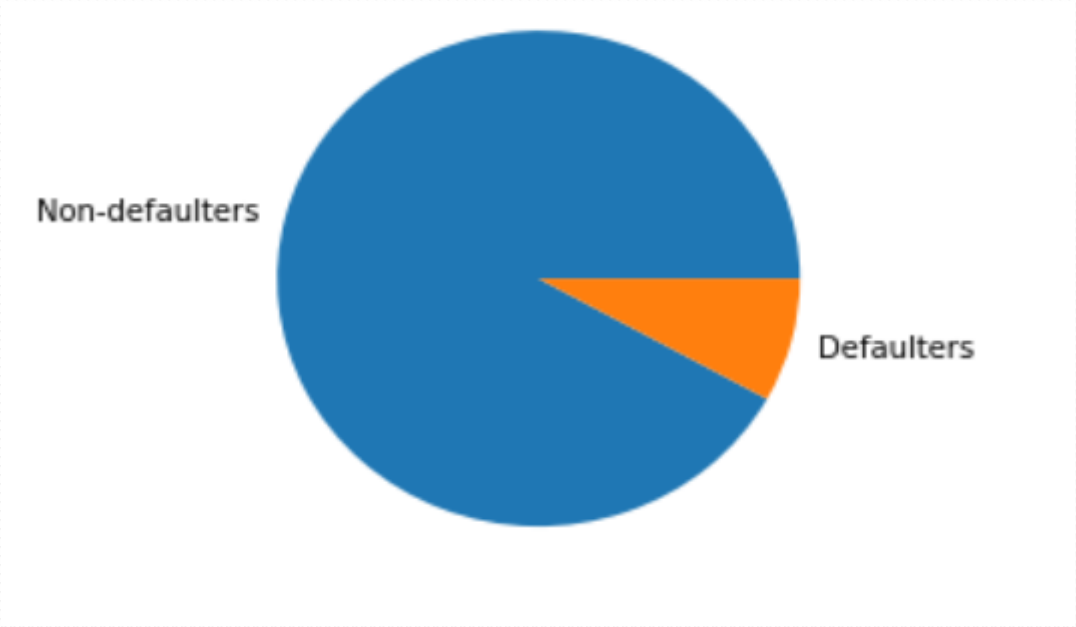

2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
CNT_FAM_MEMBERS

- Columns which are having no outliers are



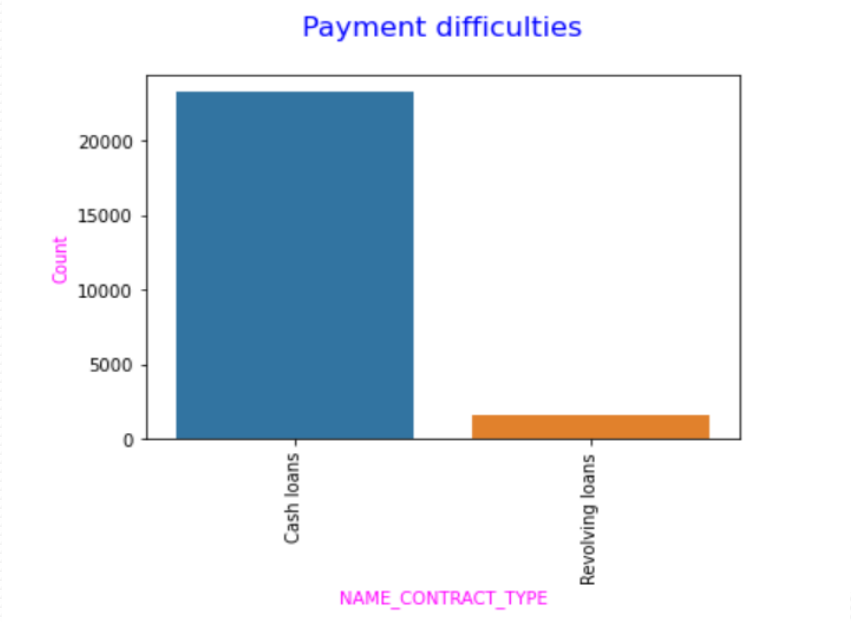
Check for Data Imbalance

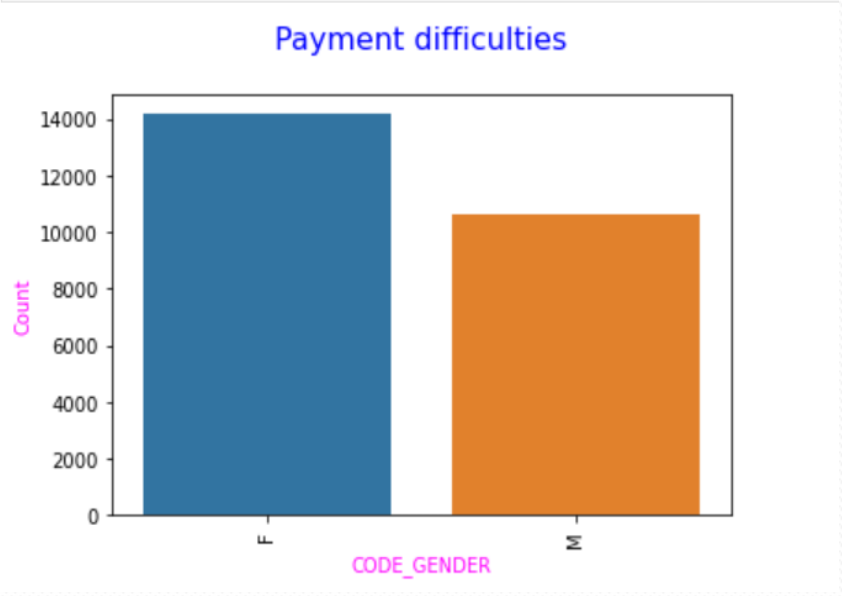
- The data is highly imbalanced the amount of applicants who have no payment difficulty (Non-defaulters) so taken only the applicants who has payment difficulties (Defaulters).



Univariate Analysis

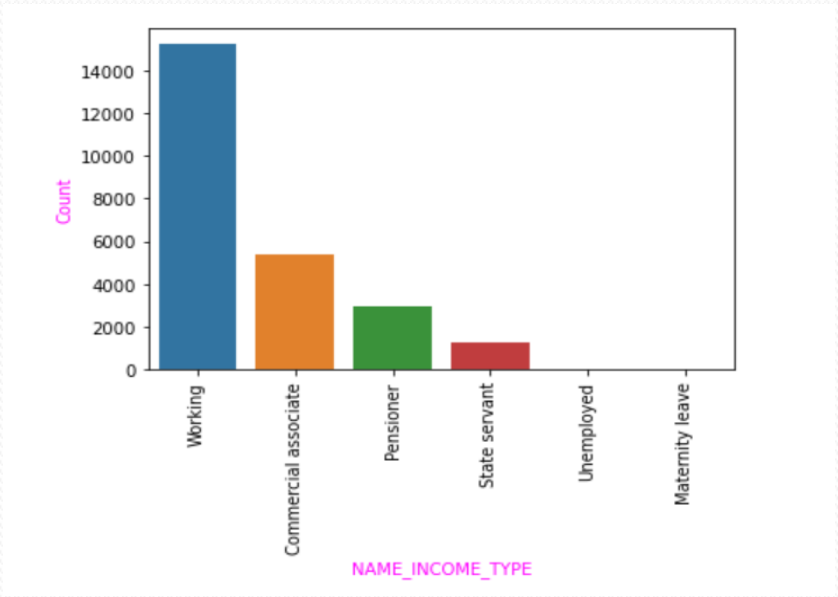
- **For Categorical Columns**
- In this columns there is prominent relationship is discovered for the defaulters.

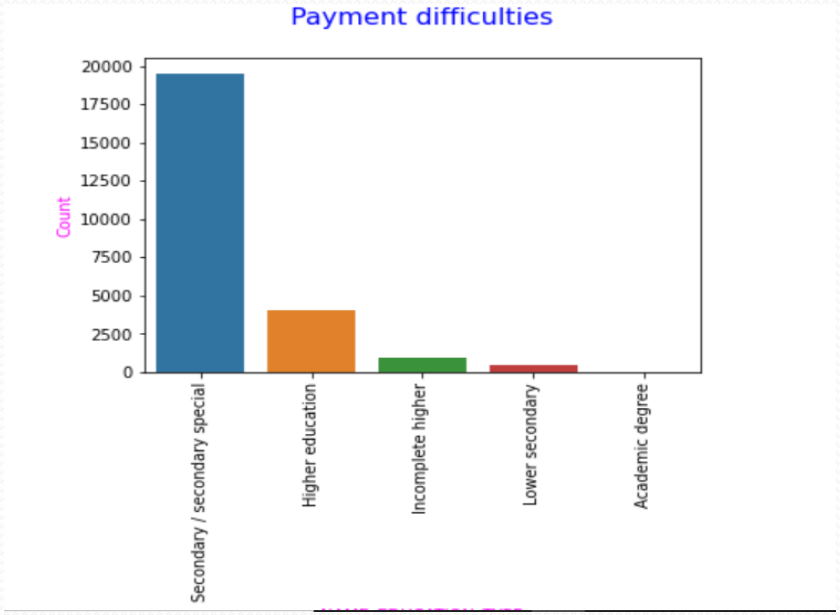


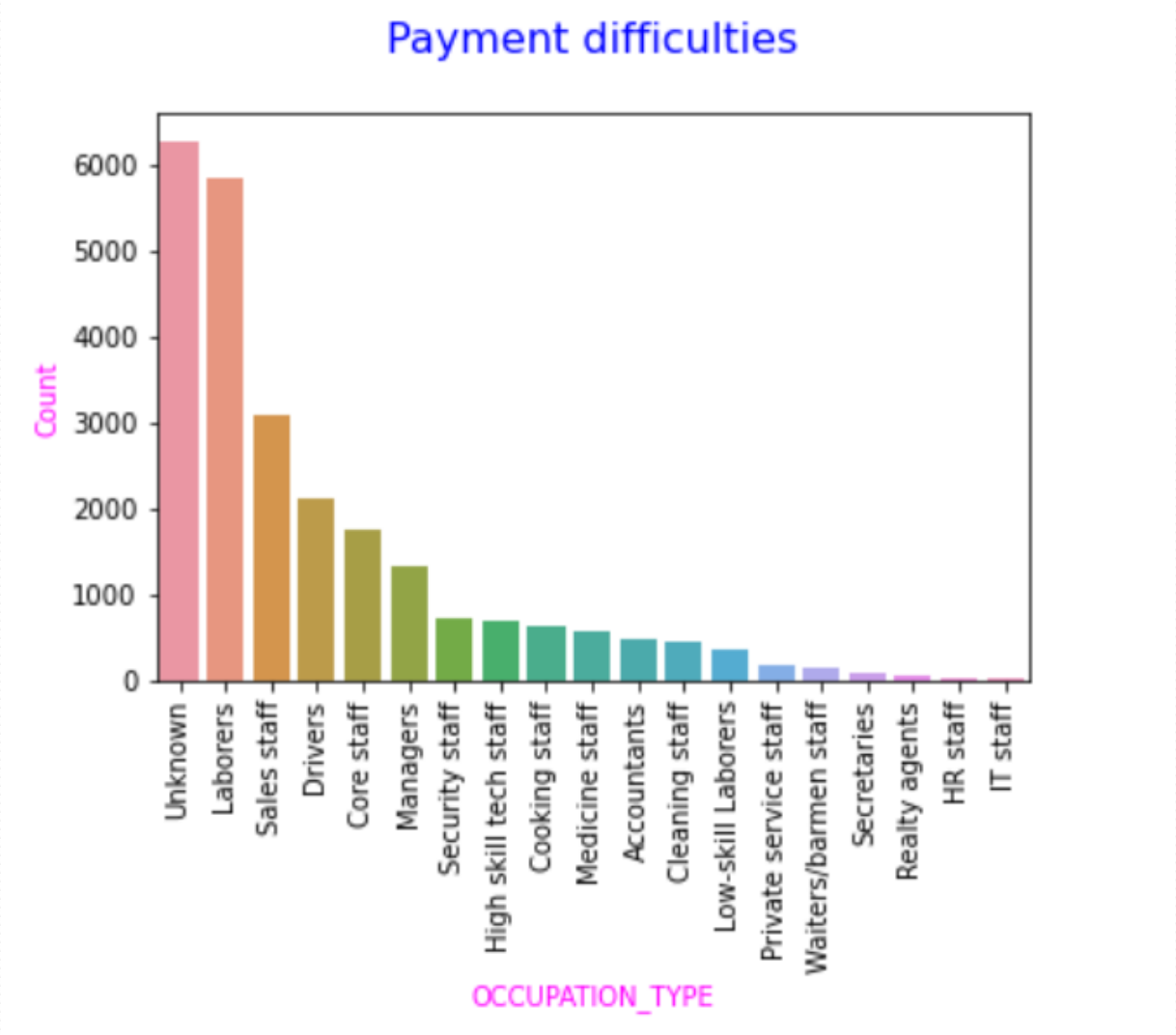



Payment difficulties
Count
17500
15000
12500
10000
7500
5000
2500
0
N
Y
FLAG_OWN_CAR

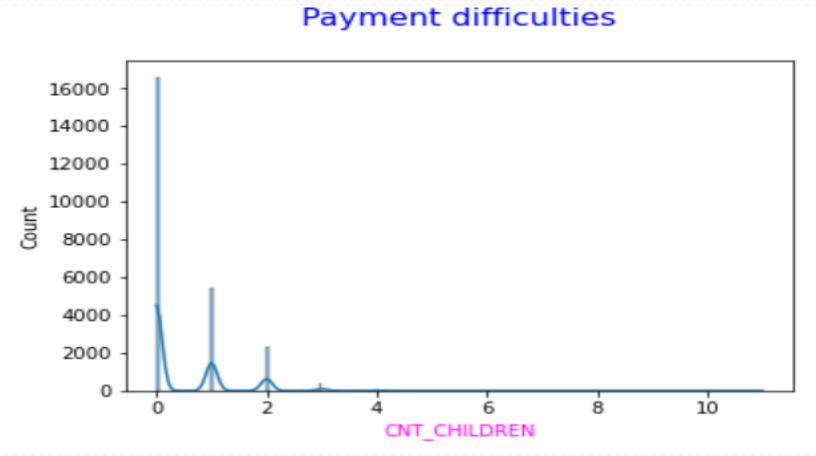

Payment difficulties
Count
16000
14000
12000
10000
8000
6000
4000
2000
0
Y
N
FLAG_OWN_REALTY

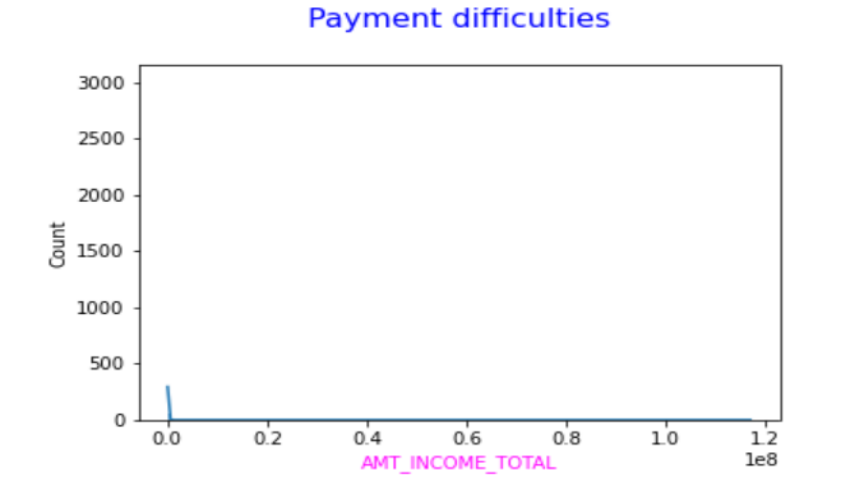

Count
14000
12000
10000
8000
6000
4000
2000
0
Working
Commercial associate
Pensioner
State servant
Unemployed
Maternity leave
NAME_INCOME_TYPE

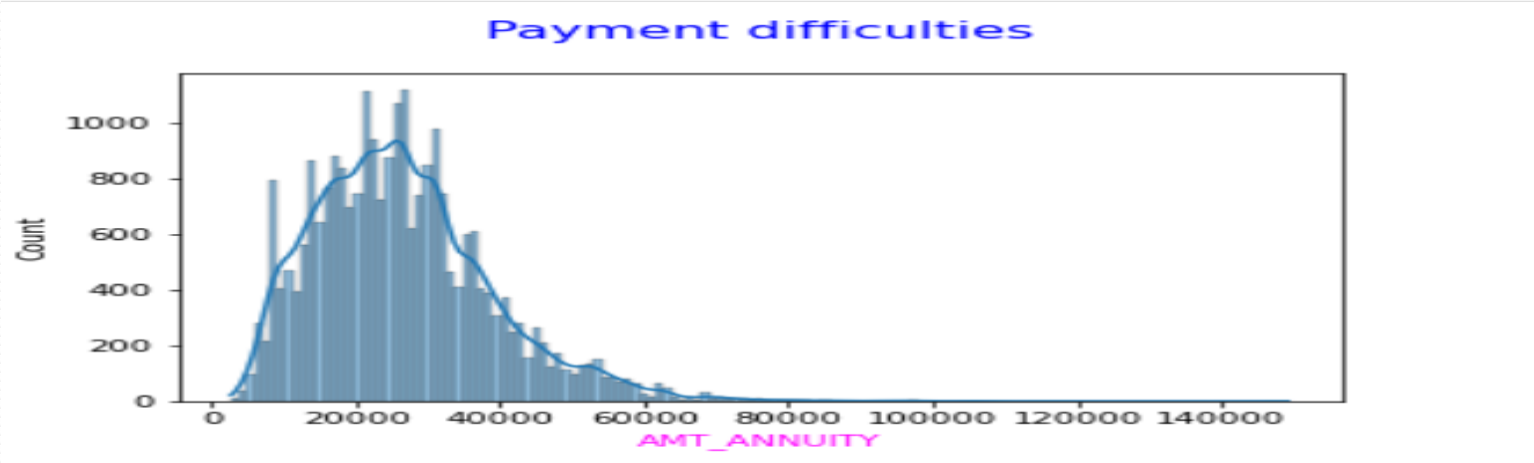

Payment difficulties
Count
20000
17500
15000
12500
10000
7500
5000
2500
0
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree

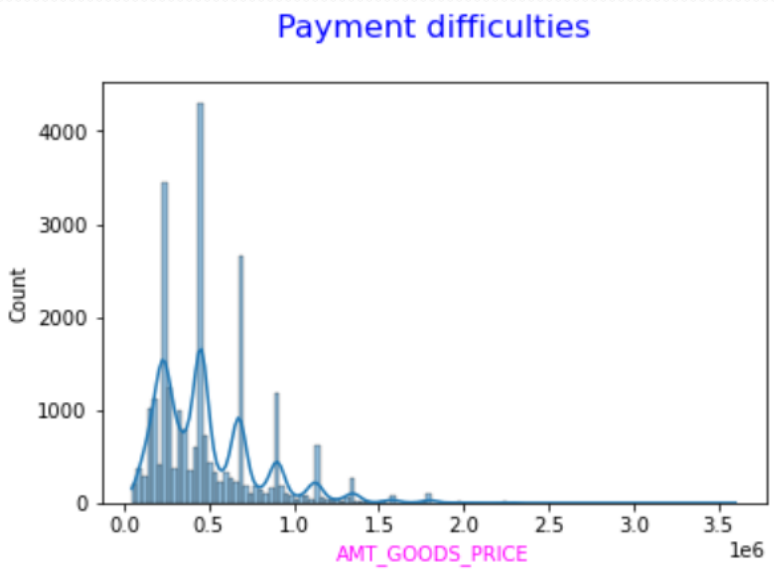

Payment difficulties
Count
6000
5000
4000
3000
2000
1000
0
Unknown
Laborers
Sales staff
Drivers
Core staff
Managers
Security staff
High skill tech staff
Cooking staff
Medicine staff
Accountants
Cleaning staff
Low-skill Laborers
Private service staff
Waiters/barmen staff
Secretaries
Realty agents
HR staff
IT staff
OCCUPATION_TYPE

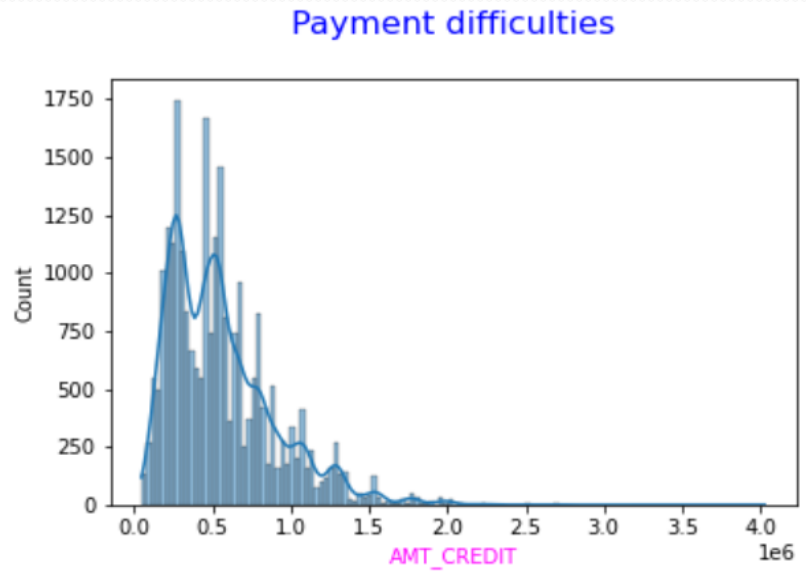
- **For Numerical Columns**
- In this some of the columns where specific relationship is present.

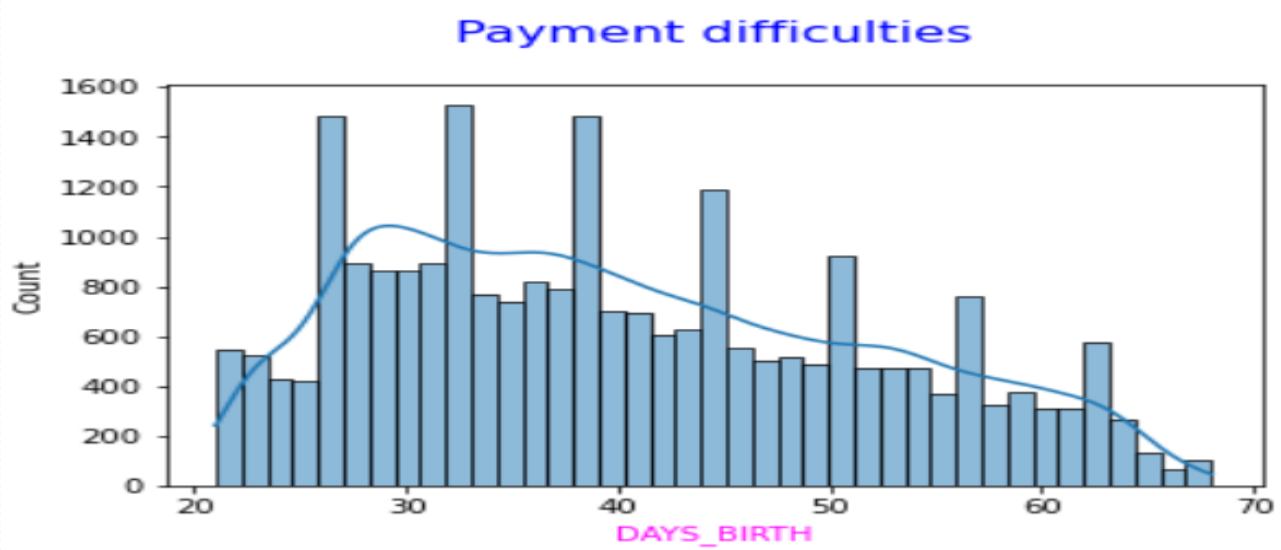






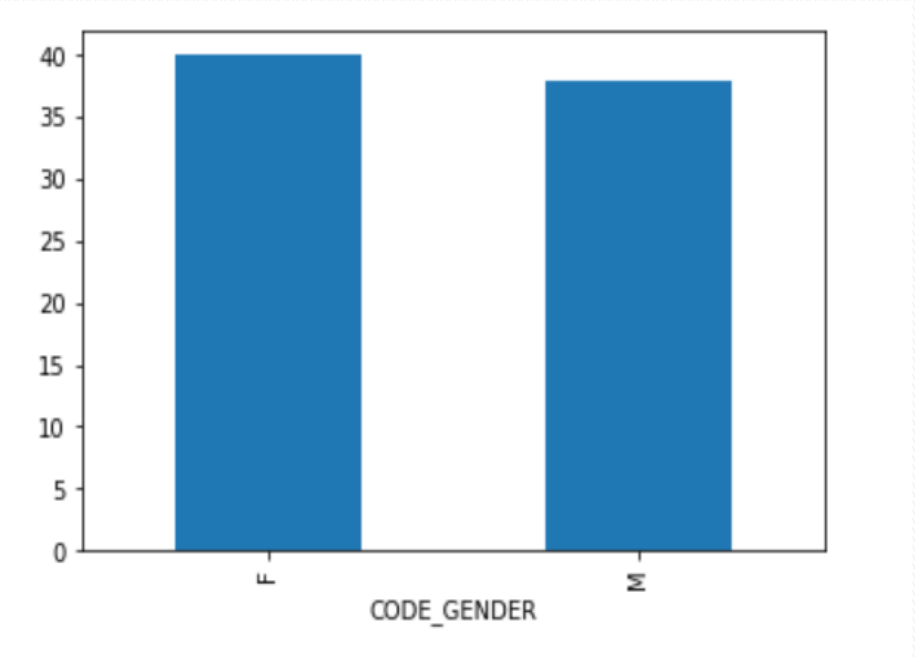

Payment difficulties
Count
AMT_GOODS_PRICE
1e6

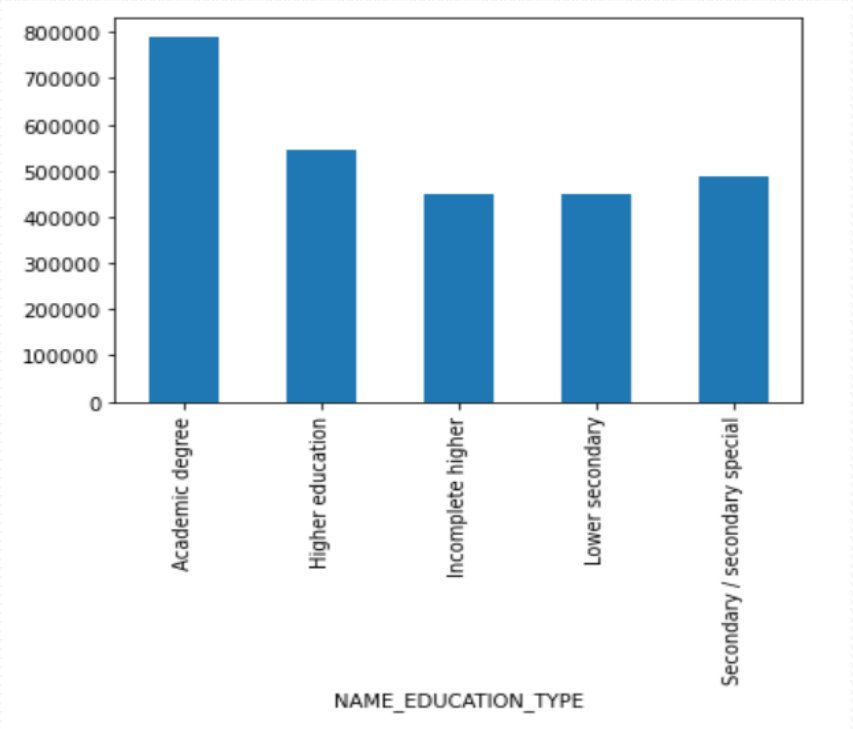

Payment difficulties
Count
AMT_CREDIT
1e6

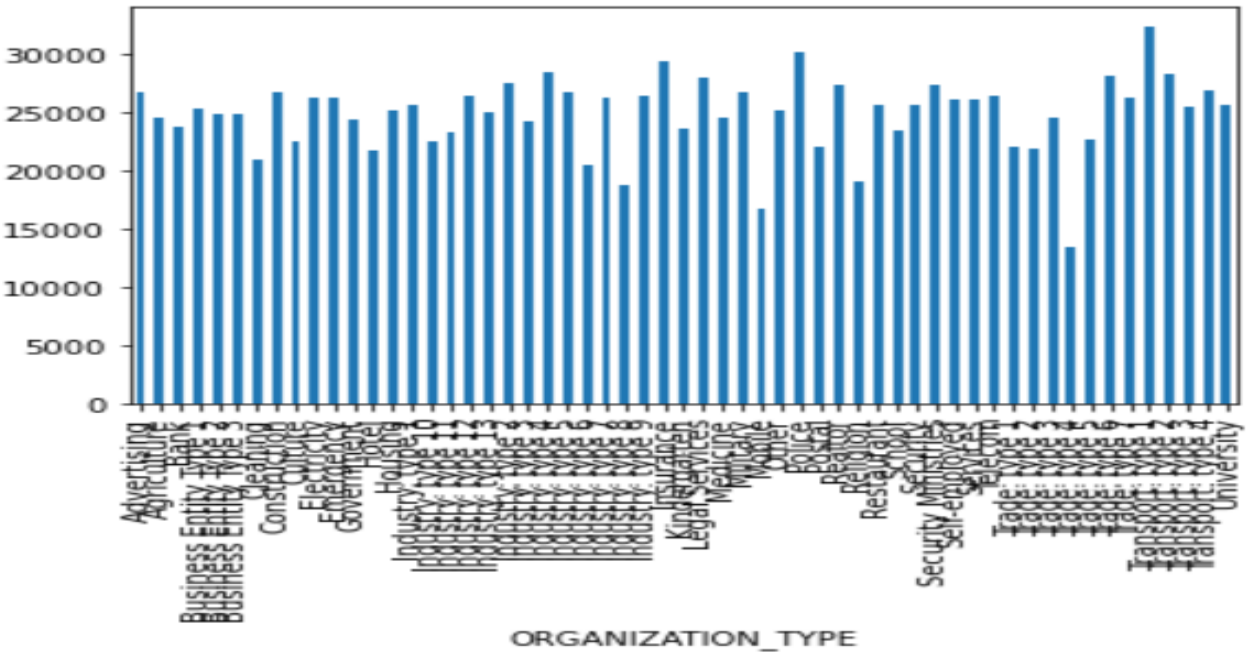

Payment difficulties
Count
DAYS_BIRTH

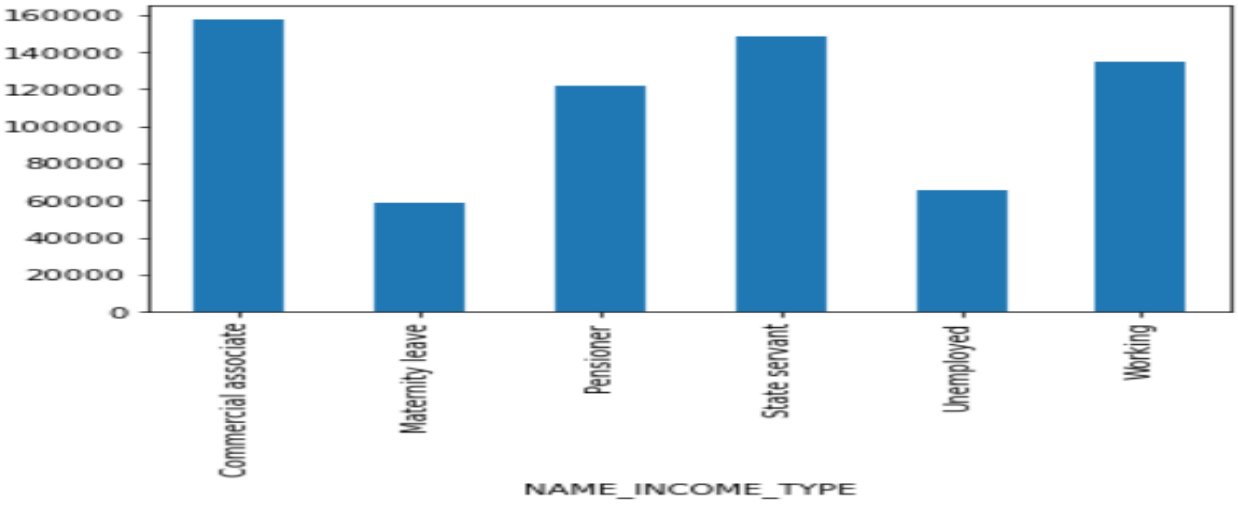
Bivariate Analysis

- **For Categorical vs Numerical Columns**
- There is mixed type of results in this analysis

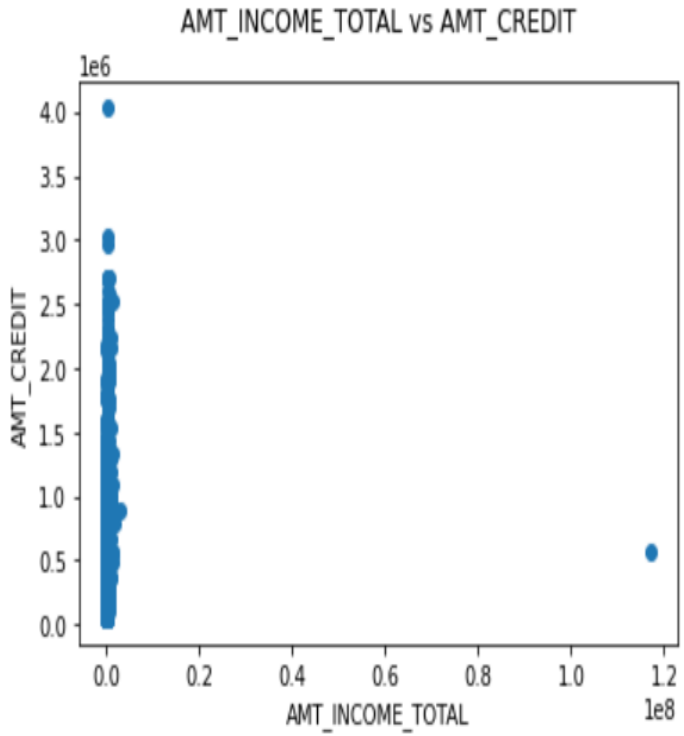


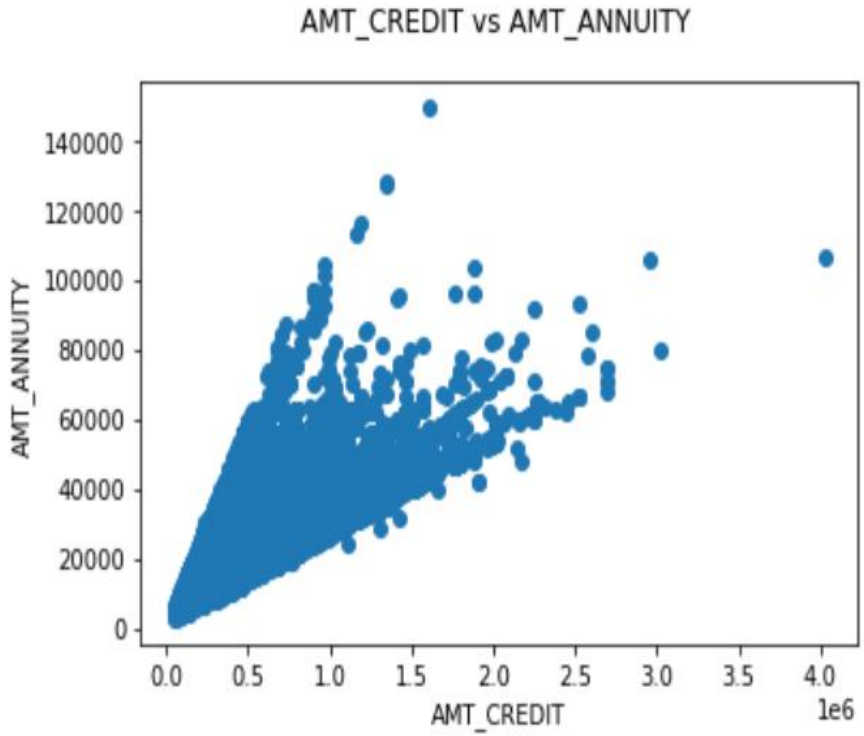


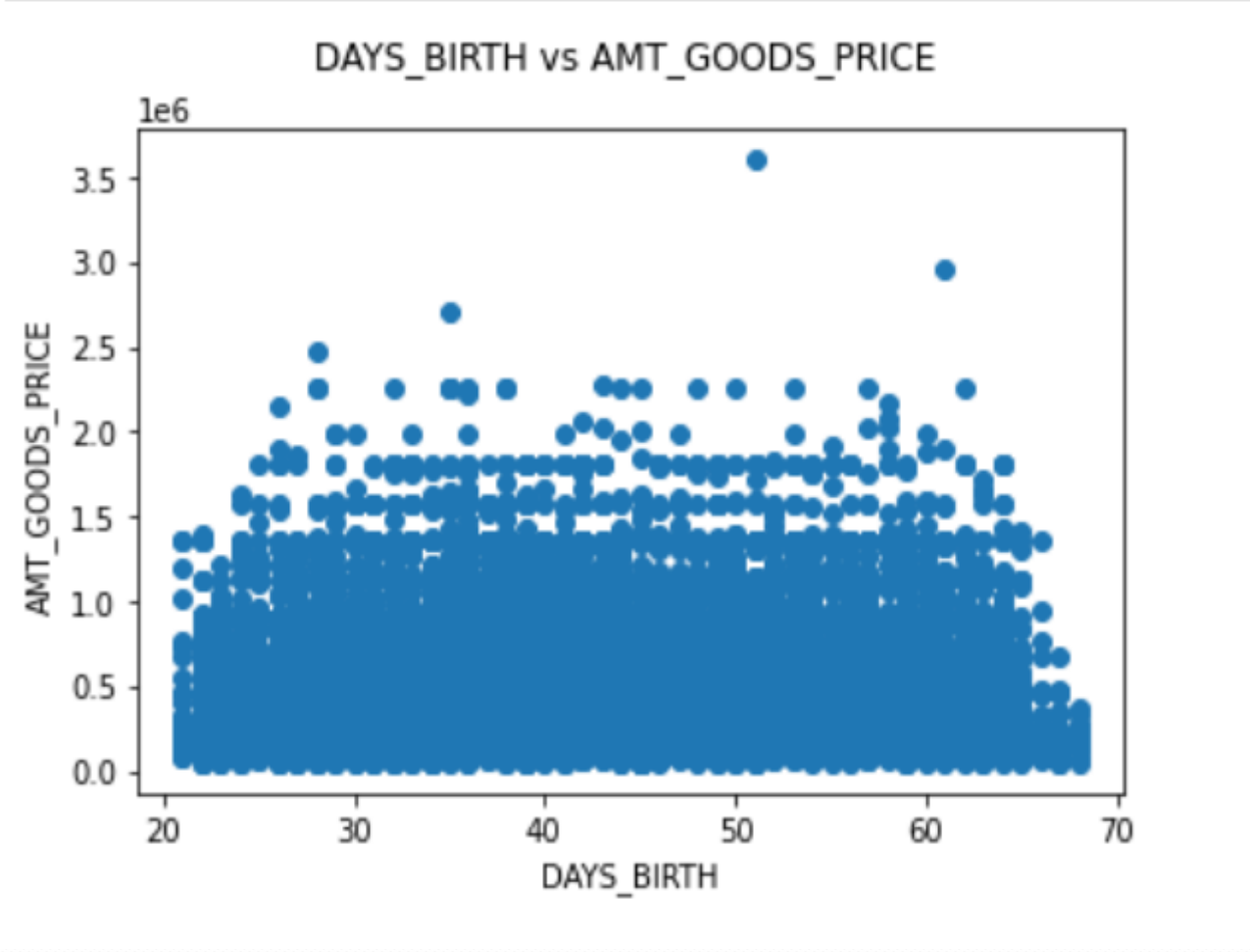

30000
25000
20000
15000
10000
5000
0
ORGANIZATION_TYPE


160000
140000
120000
100000
80000
60000
40000
20000
0
Commercial associate
Maternity leave
Pensioner
State servant
Unemployed
Working
NAME_INCOME_TYPE

- **For Numerical vs Numerical Columns**
- The most of the data present for the defaulters are in lower to middle values of the columns.

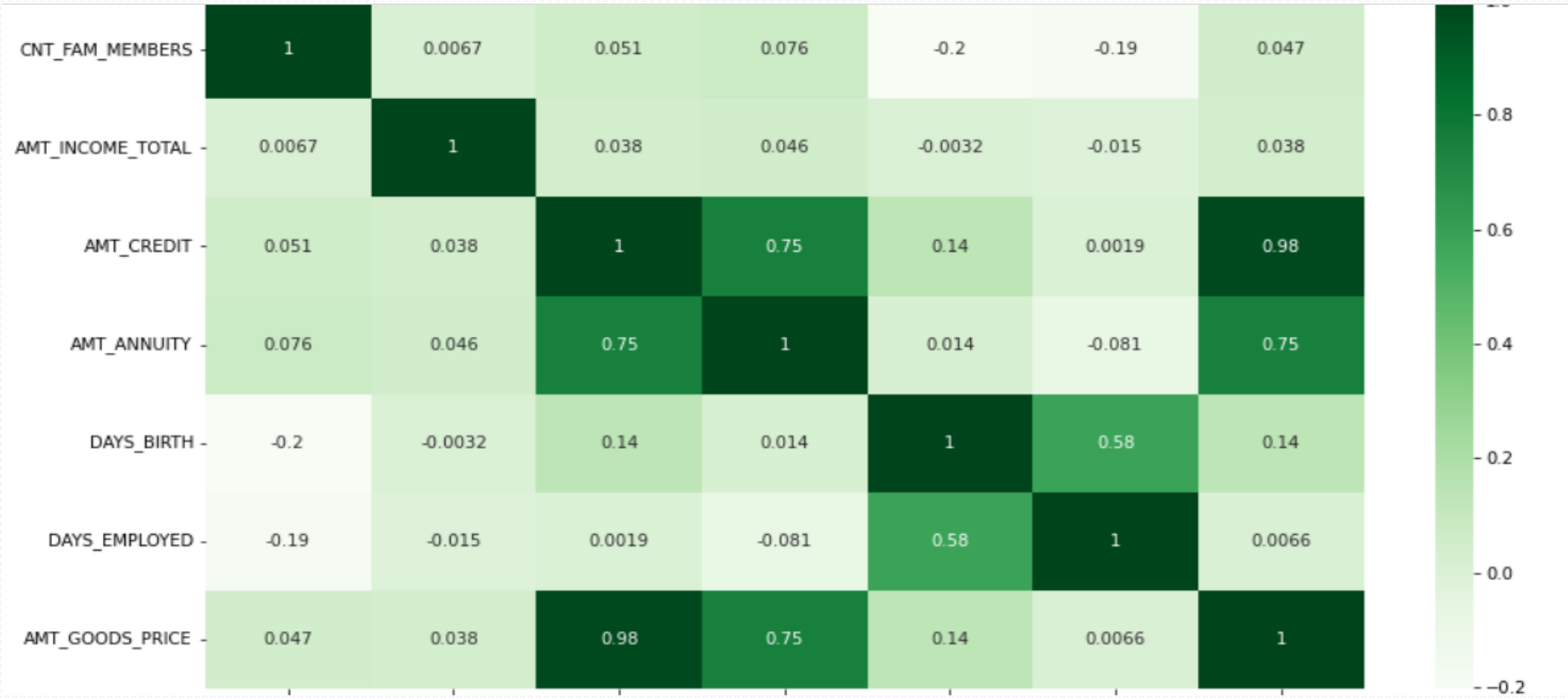





DAYS_BIRTH vs AMT_GOODS_PRICE
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
AMT_GOODS_PRICE
20
30
40
50
60
70
DAYS_BIRTH

Heatmap

- The correlation is taken for some numerical columns and there is strong realtionship present between two pairs of columns.



Insights from the data

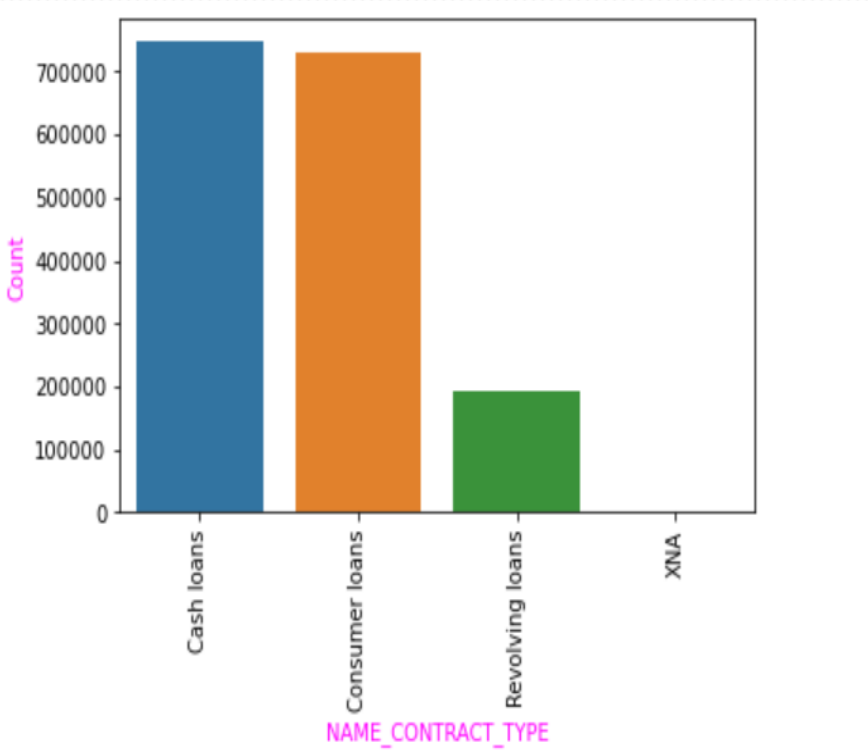
- **for target==1(having payment difficulty) only**
- Maximum applicants taken loan as cash loan and minimum have taken revolving loan.
- Females are more than Male
- Applicants having no car having high payment difficulty than having a car.
- Applicants having house or flat having high payment difficulty than no house/flat.
- Applicants having which are working has high payment difficulty and which have academic degree are having no difficulty.
- Applicants which are labour or unknown has high payment difficulty and which are IT staff are having very less difficulty.
- Maximum applicants having no children are having more payment difficulty than having children.

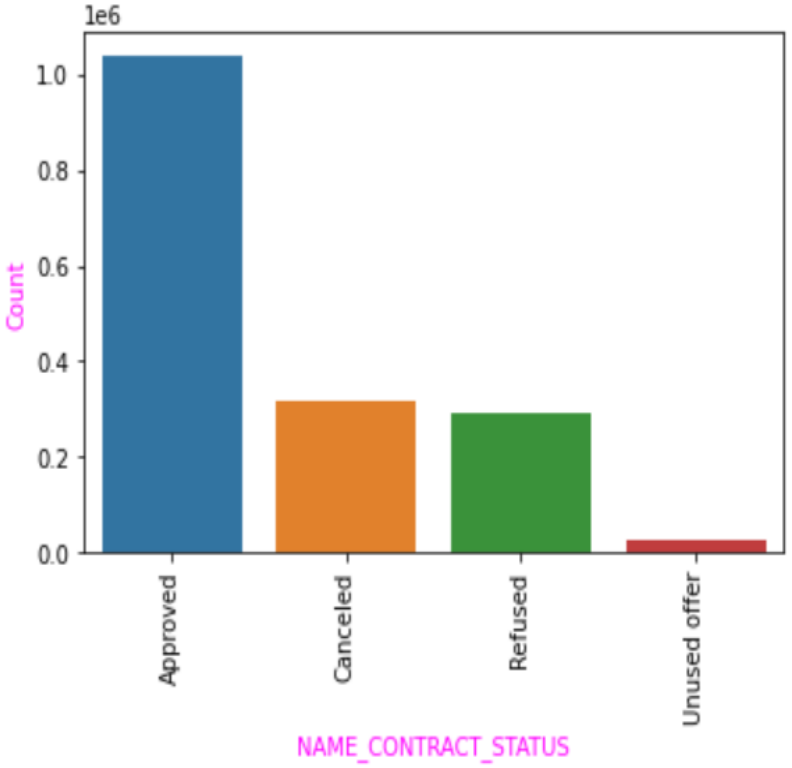
For prev_application.csv data Data Cleaning

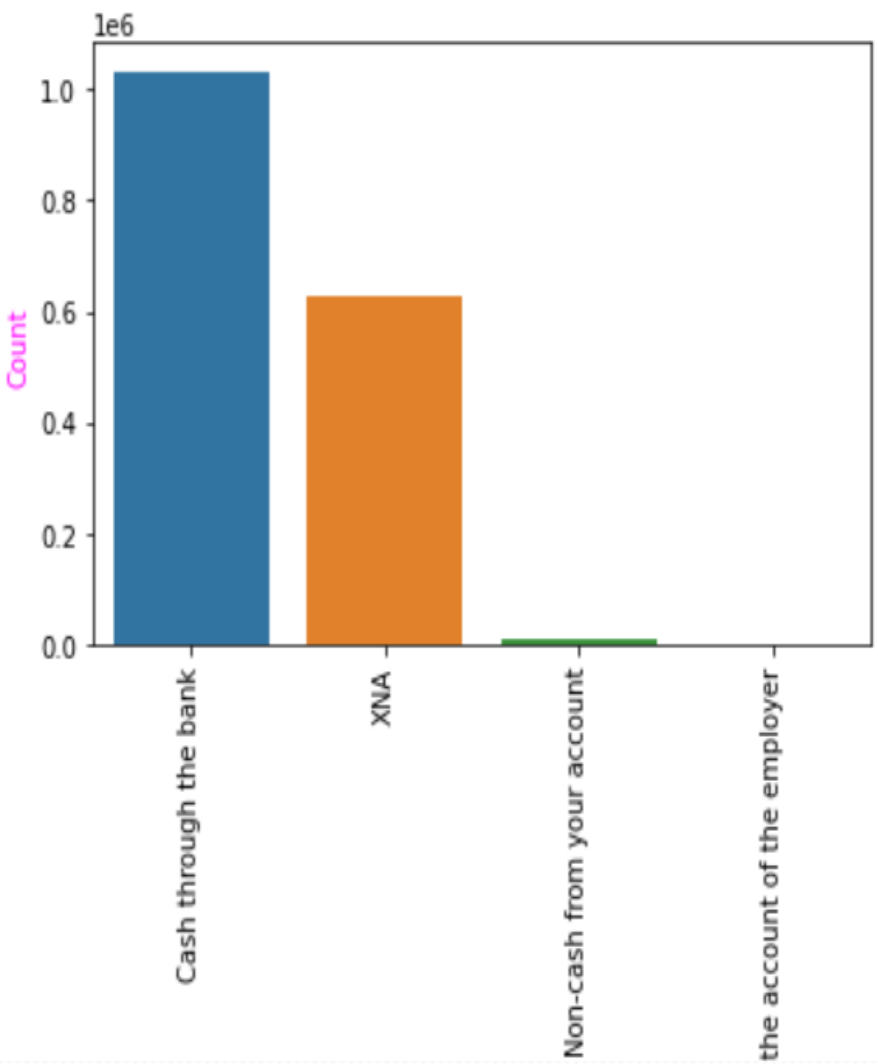
- Firtsly the shape and info about the columns is checked.
- After that columns having missing values more than 40% are dropped.
- After that missing values are imputed.

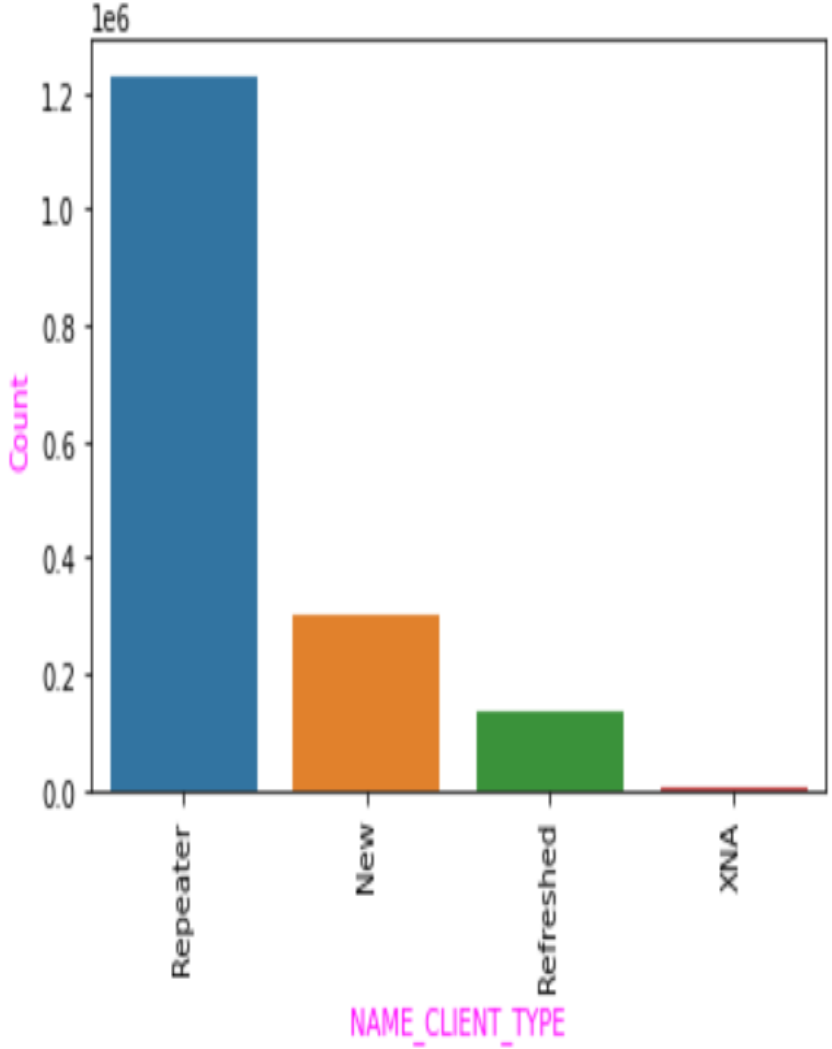
Univariate Analysis

- **For Categorical Columns**

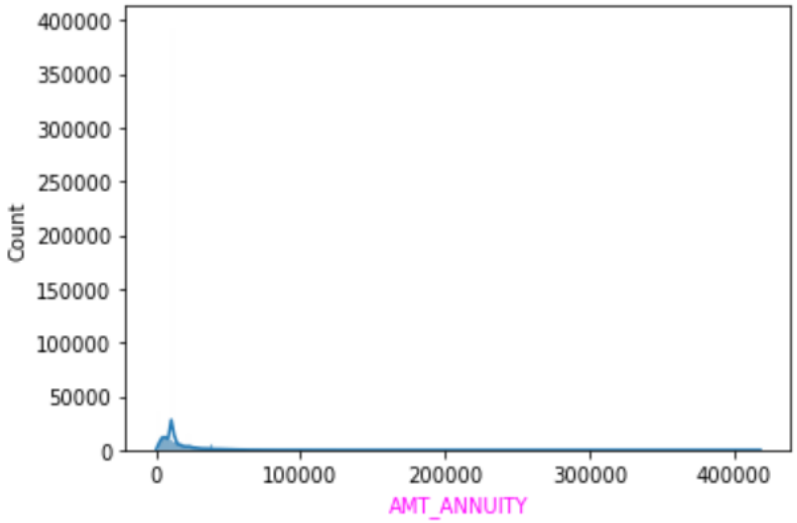


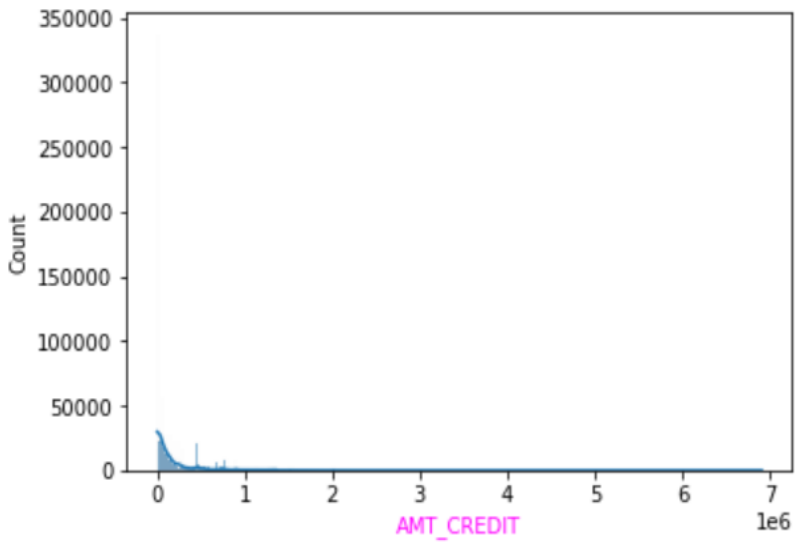


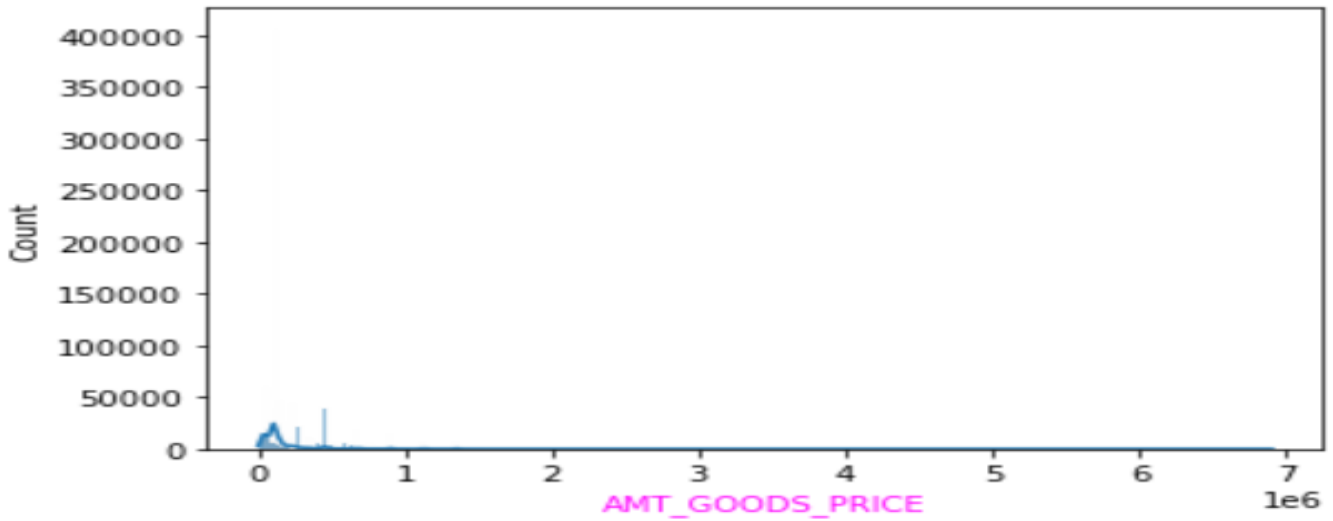

1e6
1.0
0.8
0.6
0.4
0.2
0.0
Count
Cash through the bank
XNA
Non-cash from your account
the account of the employer


1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Count
Repeater
New
Refreshed
XNA
NAME_CLIENT_TYPE

- **For Numerical Columns**

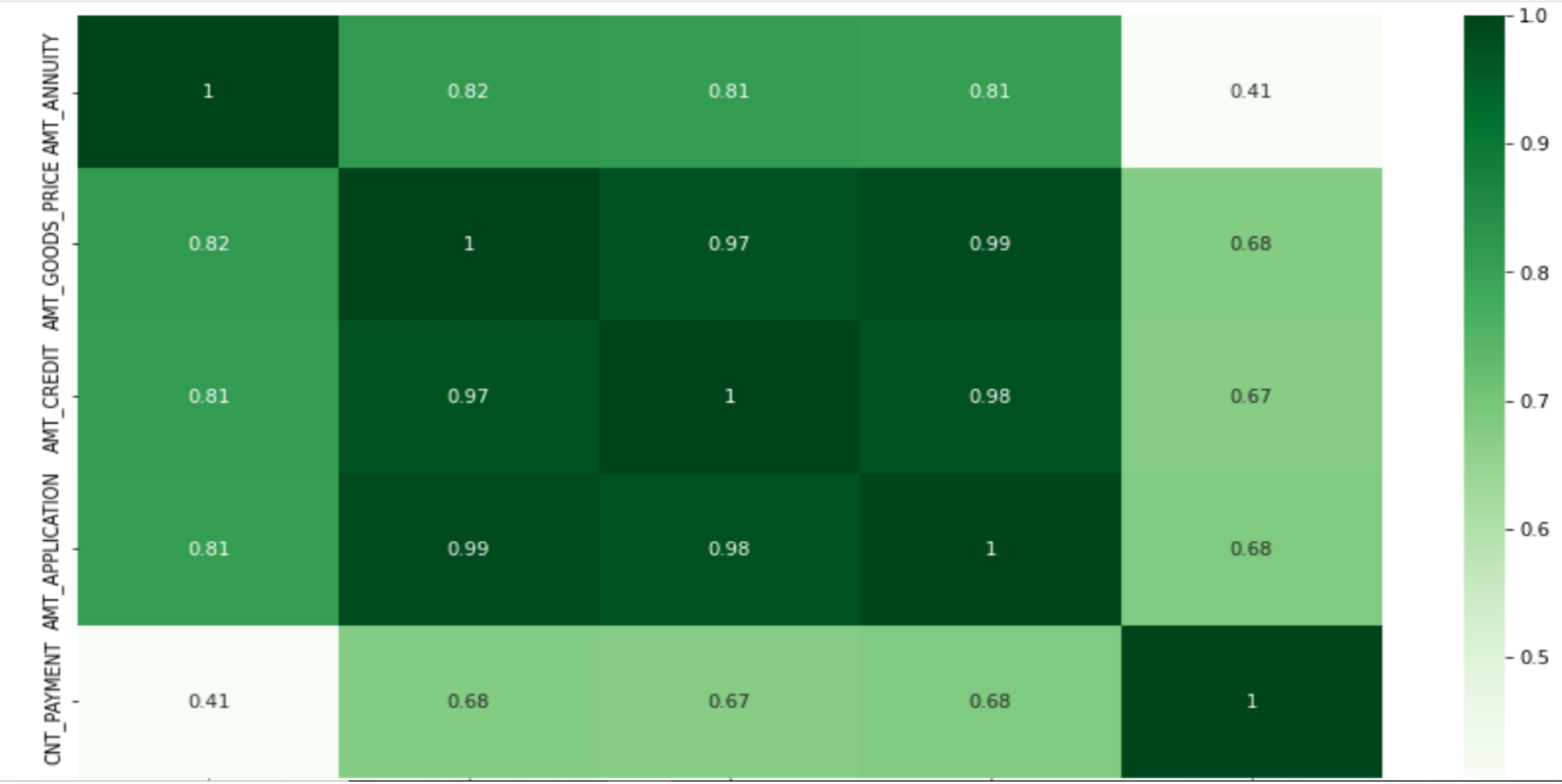






Heatmap

- In this heatmap the most of the columns had strong relationship between them.



Insights from the data

- Maximum of applicants taken cash loans or consumer loan and minimum applicants taken revolving loan.
- Around 70% of applicants has their loan approved and 30% of applicants have their loan refused or canceled.
- Maximum applicats taken their loan as cash through the bank and and very minimum have taken cashless.
- 80% applicants are repeaters and 20% are refreshed or new.
- Maximum applicants taken credit less than 100000.
- maximum applicants taken annuity less than 30000.

Thank You