Credit EDA Case Study

Prepared by – Indranil Kundu & Divya Dindorkar

Problem Statement

This case study aims to identify patterns which indicate if a client has difficulties 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.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Structure of Application_data

- Dimension of Application Dataframe: (307511, 122)
- Columns in Dataframe which have maximum % of missing data [>= 40%]:

COMMONAREA_MEDI	69.87
COMMONAREA_AVG	69.87
COMMONAREA_MODE	69.87
NONLIVINGAPARTMENTS_MODE	69.43
NONLIVINGAPARTMENTS_MEDI	69.43
NONLIVINGAPARTMENTS_AVG	69.43
FONDKAPREMONT_MODE	68.39
LIVINGAPARTMENTS_MEDI	68.35
LIVINGAPARTMENTS_MODE	68.35
LIVINGAPARTMENTS_AVG	68.35
FLOORSMIN_MEDI	67.85
FLOORSMIN_MODE	67.85
FLOORSMIN_AVG	67.85
YEARS_BUILD_MEDI	66.50
YEARS_BUILD_AVG	66.50
YEARS_BUILD_MODE	66.50
OWN_CAR_AGE	65.99
LANDAREA_MODE	59.38
LANDAREA_AVG	59.38
LANDAREA_MEDI	59.38
BASEMENTAREA_MEDI	58.52
BASEMENTAREA_AVG	58.52
BASEMENTAREA_MODE	58.52
EXT_SOURCE_1	56.38
NONLIVINGAREA_MEDI	55.18
NONLIVINGAREA_AVG	55.18
NONLIVINGAREA_MODE	55.18
ELEVATORS_MODE	53.30
ELEVATORS_AVG	53.30
ELEVATORS_MEDI	53.30
WALLSMATERIAL_MODE	50.84
APARTMENTS_MODE	50.75
APARTMENTS_AVG	50.75
APARTMENTS_MEDI	50.75
ENTRANCES_MEDI	50.35
ENTRANCES_MODE	50.35
ENTRANCES_AVG	50.35
LIVINGAREA_MEDI	50.19
LIVINGAREA_MODE	50.19
LIVINGAREA_AVG	50.19
HOUSETYPE_MODE	50.18
FLOORSMAX_MODE	49.76
FLOORSMAX_MEDI	49.76
FLOORSMAX_AVG	49.76
YEARS_BEGINEXPLUATATION_MEDI	48.78
YEARS_BEGINEXPLUATATION_AVG	48.78
YEARS_BEGINEXPLUATATION_MODE	48.78
TOTALAREA_MODE	48.27
EMERGENCYSTATE_MODE	47.40

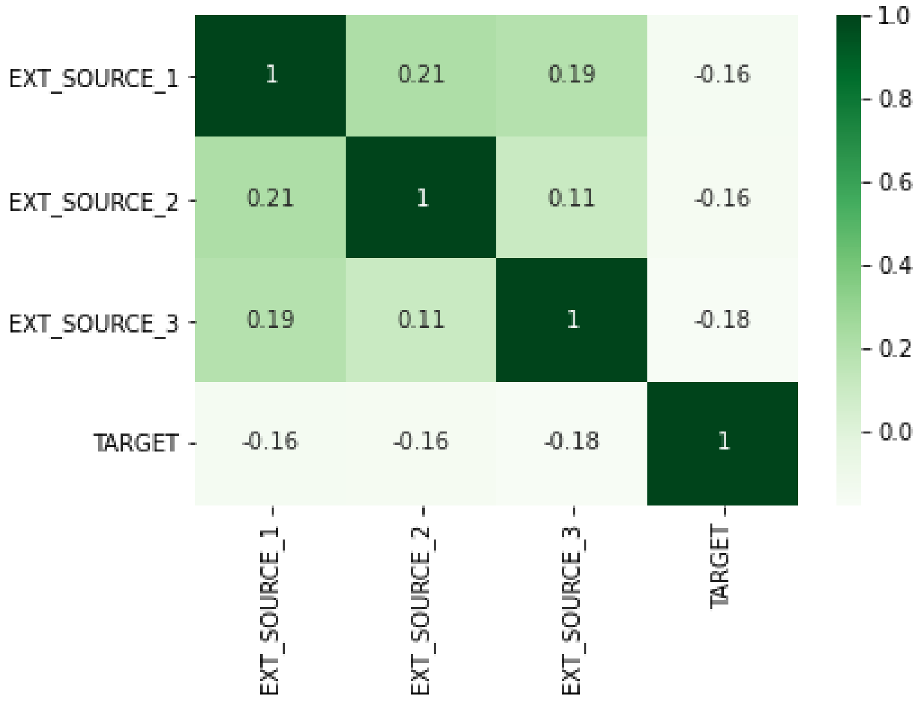
Insight: There are 49 columns which have more than or equal to 40% missing values. Which can be removed.

Analyze and Delete Unnecessary Columns

Analyzing and deleting columns which are not necessary for analysis

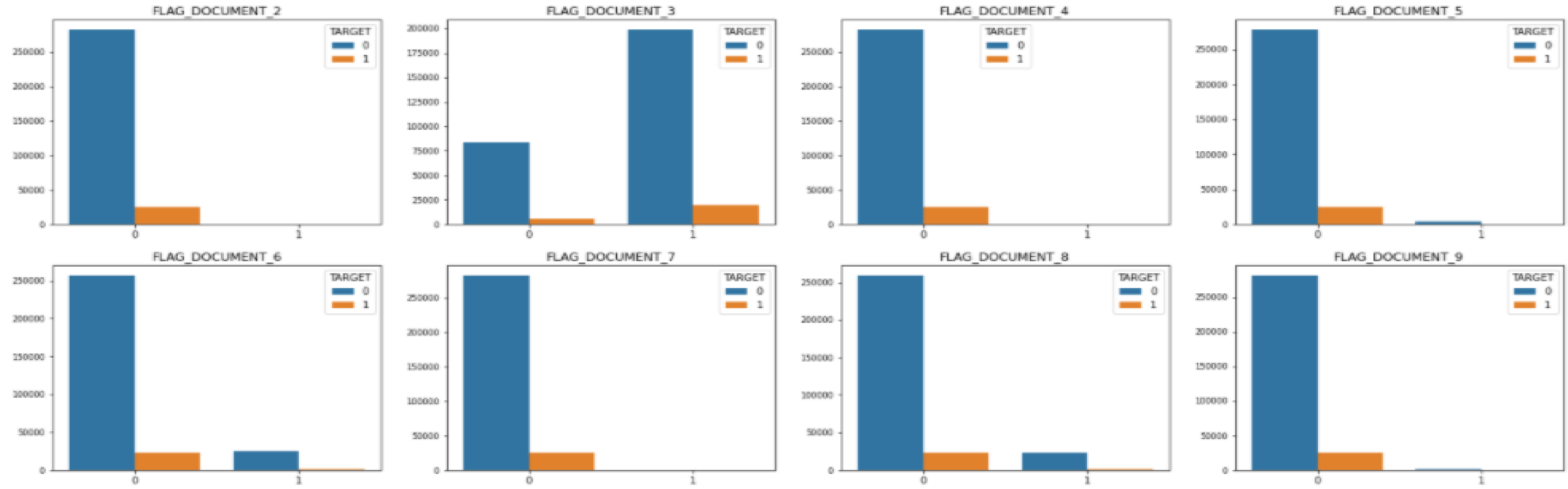
'EXT_SOURCE_1','EXT_SOURCE_2','EXT_SOURCE_3‘ vs. ‘TARGET’

- 'EXT_SOURCE_1','EXT_SOURCE_2','EXT_SOURCE_3' and 'TARGET' column

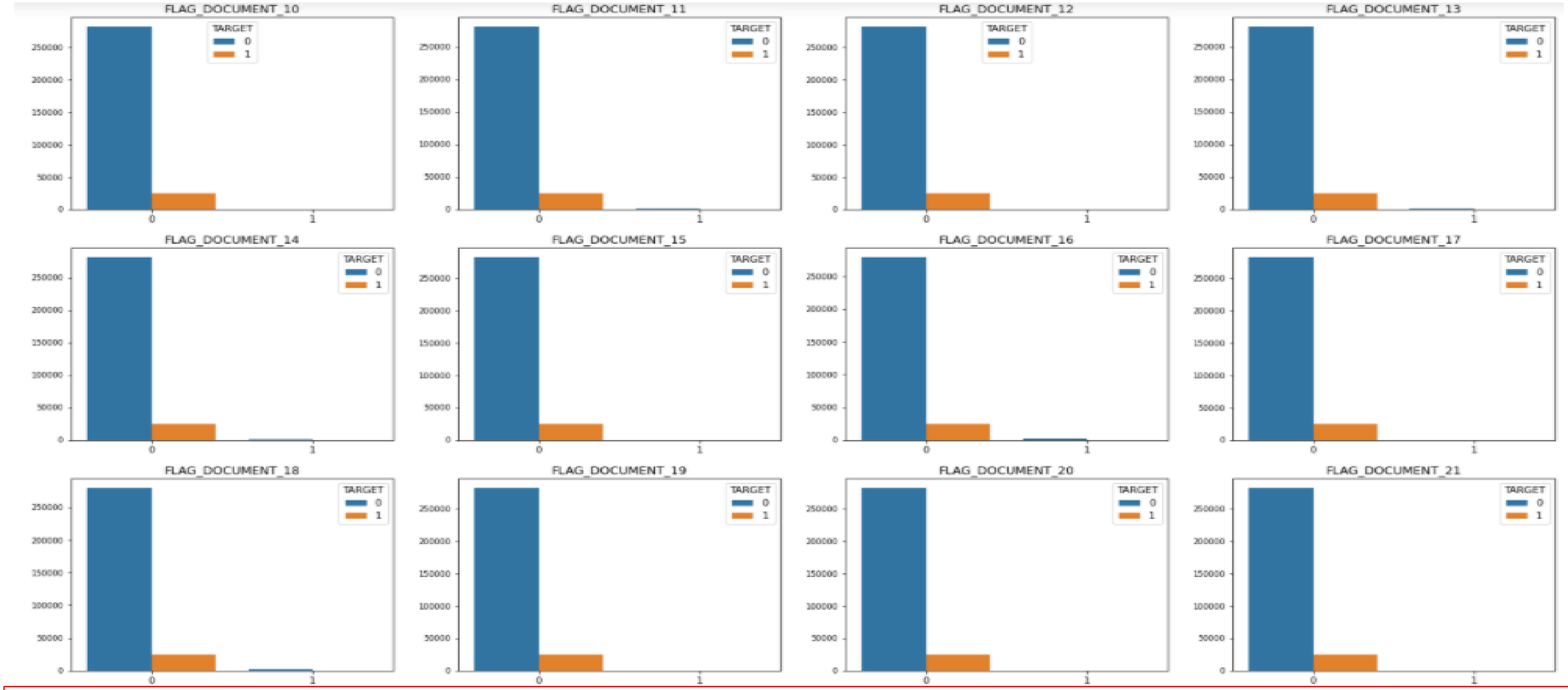


Insight:
There is no correlation between 'EXT_SOURCE_1','EXT_SOURCE_2','EXT_SOURCE_3' and 'TARGET‘ columns. We can thus drop these columns.

Analysis of ‘FLAG_DOCUMENTS’ Columns

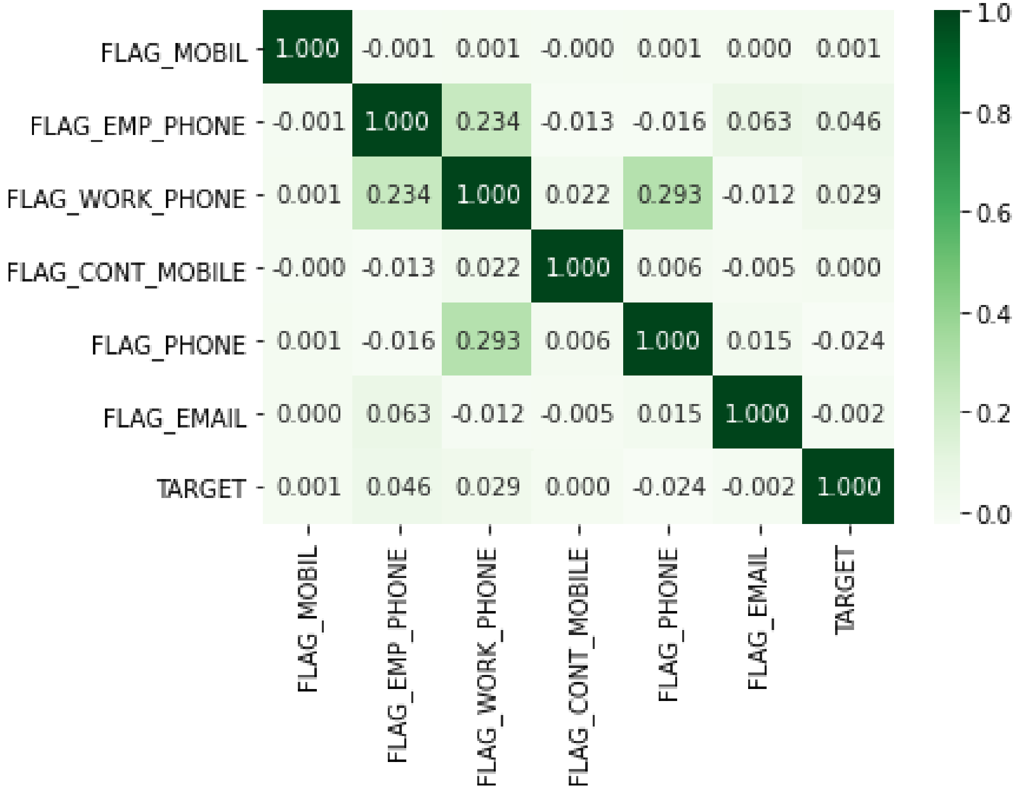


Cont…



Inference:
From the above graph we can infer that in most of the loan application cases, people who submitted 'FLAG_DOCUMENT_3' had a less chance of defaulting . So, we can keep this column and delete all the other columns.

Contact Based Columns vs. Target



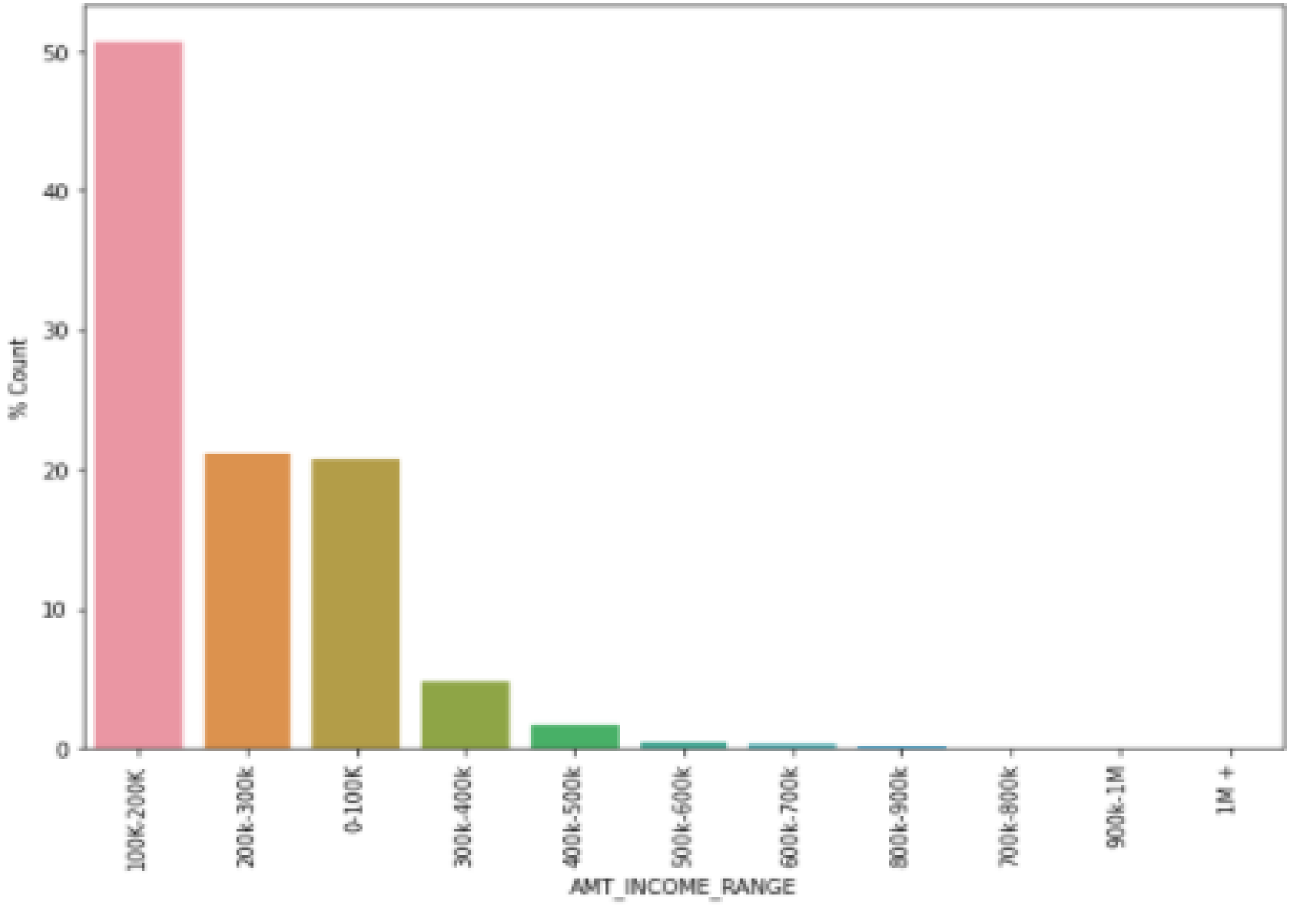
Inference:
No correlation between the analyzed columns with 'TARGET'. We can thus drop these columns.

Structure Post Dropping Columns

- Post dropping unwanted data, the current data dimensions are : (307511, 46)
- Below are the columns left:

#	Column	Non-Null Count	Dtype
0	SK_ID_CURR	307511 non-null	int64
1	TARGET	307511 non-null	int64
2	NAME_CONTRACT_TYPE	307511 non-null	object
3	CODE_GENDER	307511 non-null	object
4	FLAG_OWN_CAR	307511 non-null	object
5	FLAG_OWN_REALTY	307511 non-null	object
6	CNT_CHILDREN	307511 non-null	int64
7	AMT_INCOME_TOTAL	307511 non-null	float64
8	AMT_CREDIT	307511 non-null	float64
9	AMT_ANNUITY	307499 non-null	float64
10	AMT_GOODS_PRICE	307233 non-null	float64
11	NAME_TYPE_SUITE	306219 non-null	object
12	NAME_INCOME_TYPE	307511 non-null	object
13	NAME_EDUCATION_TYPE	307511 non-null	object
14	NAME_FAMILY_STATUS	307511 non-null	object
15	NAME_HOUSING_TYPE	307511 non-null	object
16	REGION_POPULATION_RELATIVE	307511 non-null	float64
17	DAYS_BIRTH	307511 non-null	int64
18	DAYS_EMPLOYED	307511 non-null	int64
19	DAYS_REGISTRATION	307511 non-null	float64
20	DAYS_ID_PUBLISH	307511 non-null	int64
21	OCCUPATION_TYPE	211120 non-null	object
22	CNT_FAM_MEMBERS	307509 non-null	float64
23	REGION_RATING_CLIENT	307511 non-null	int64
24	REGION_RATING_CLIENT_W_CITY	307511 non-null	int64
25	WEEKDAY_APPR_PROCESS_START	307511 non-null	object
26	HOUR_APPR_PROCESS_START	307511 non-null	int64
27	REG_REGION_NOT_LIVE_REGION	307511 non-null	int64
28	REG_REGION_NOT_WORK_REGION	307511 non-null	int64
29	LIVE_REGION_NOT_WORK_REGION	307511 non-null	int64
30	REG_CITY_NOT_LIVE_CITY	307511 non-null	int64
31	REG_CITY_NOT_WORK_CITY	307511 non-null	int64
32	LIVE_CITY_NOT_WORK_CITY	307511 non-null	int64
33	ORGANIZATION_TYPE	307511 non-null	object
34	OBS_30_CNT_SOCIAL_CIRCLE	306490 non-null	float64
35	DEF_30_CNT_SOCIAL_CIRCLE	306490 non-null	float64
36	OBS_60_CNT_SOCIAL_CIRCLE	306490 non-null	float64
37	DEF_60_CNT_SOCIAL_CIRCLE	306490 non-null	float64
38	DAYS_LAST_PHONE_CHANGE	307510 non-null	float64
39	FLAG_DOCUMENT_3	307511 non-null	int64
40	AMT_REQ_CREDIT_BUREAU_HOUR	265992 non-null	float64
41	AMT_REQ_CREDIT_BUREAU_DAY	265992 non-null	float64
42	AMT_REQ_CREDIT_BUREAU_WEEK	265992 non-null	float64
43	AMT_REQ_CREDIT_BUREAU_MON	265992 non-null	float64
44	AMT_REQ_CREDIT_BUREAU_QRT	265992 non-null	float64
45	AMT_REQ_CREDIT_BUREAU_YEAR	265992 non-null	float64

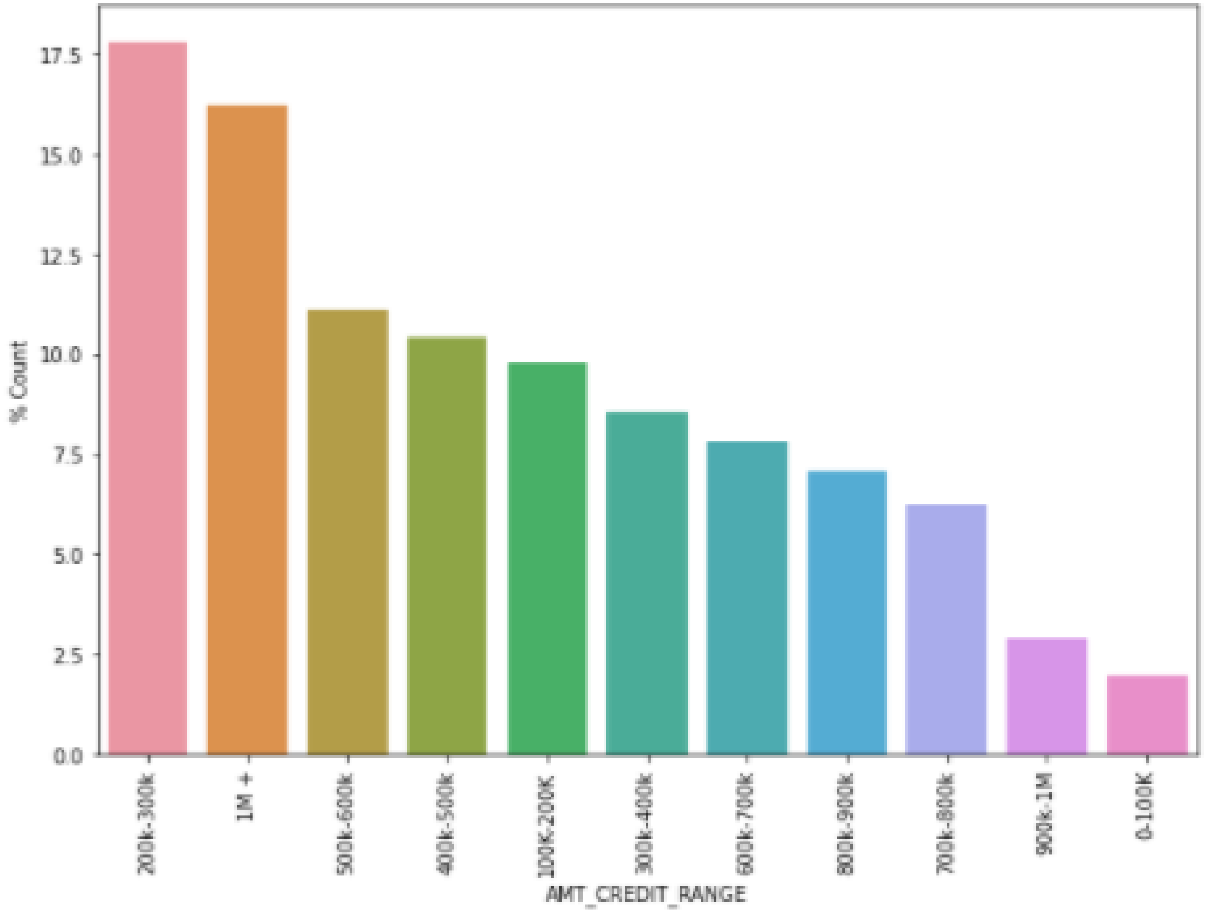
Analysis of 'AMT_INCOME_RANGE' column



100K-200K	50.735000
200k-300k	21.210691
0-100K	20.729695
300k-400k	4.776116
400k-500k	1.744669
500k-600k	0.356354
600k-700k	0.282805
800k-900k	0.096980
700k-800k	0.052721
900k-1M	0.009112
1M +	0.005858

Inference:
More than 50% applicants have Income amount in the range 100K-200K. Almost 97% Loan applicants have income less than 500k.

Analysis of ‘AMT_CREDIT’ Column

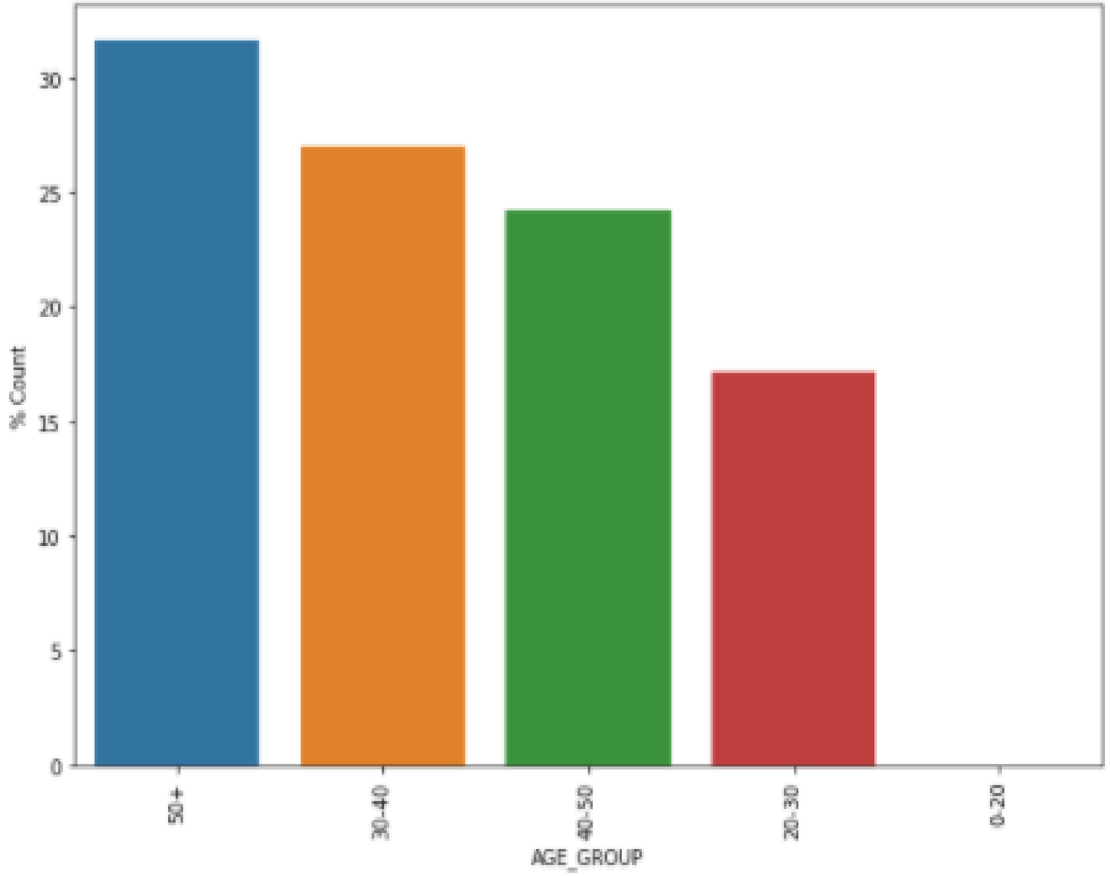



```
200k-300k     17.824728
1M +          16.254703
500k-600k     11.131960
400k-500k     10.418489
100K-200K      9.801275
300k-400k      8.564897
600k-700k      7.820533
800k-900k      7.086576
700k-800k      6.241403
900k-1M        2.902986
0-100K         1.952450
Name: AMT_CREDIT_RANGE, dtype: float64
```

Inference:
- More than 16% loan applicants have taken loan of more than 1M.
- More than 17% loan applicants have taken loan in the range 200k-300k.

Analysis of ‘AGE_GROUP’ Column

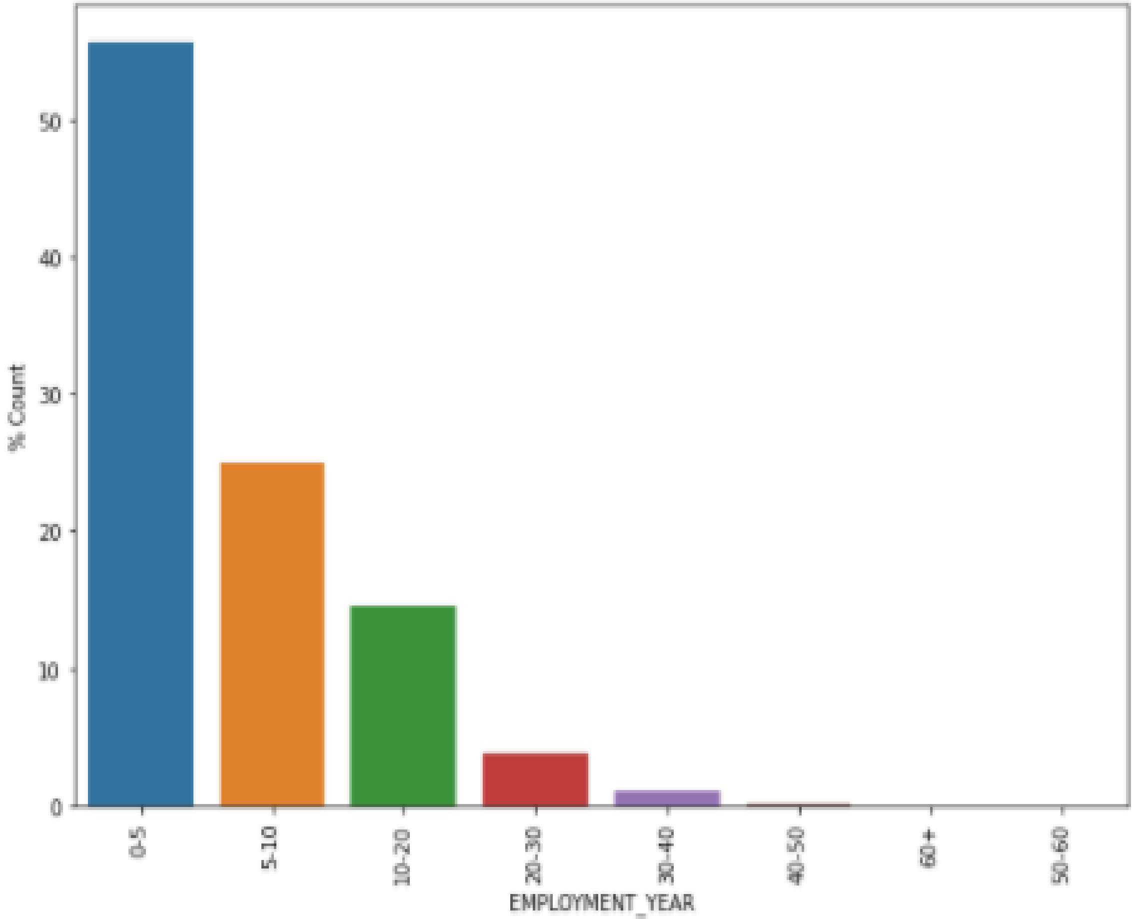
```
50+       31.604398
30-40     27.028952
40-50     24.194582
20-30     17.171743
0-20       0.000325
Name: AGE_GROUP, dtype: float64
```



Inference:
- 31% of loan applicants have age more than 50.
- 27% of loan applicants fall under the age group 30-40.

Analysis of ‘YEARS_EMPLOYED’ Column

```
0-5      55.582363
5-10     24.966441
10-20    14.564315
20-30     3.750117
30-40     1.058720
40-50     0.078044
60+       0.000000
50-60     0.000000
Name: EMPLOYMENT_YEAR, dtype: float64
```



Inference:
- 55% of loan applicants have work experience between 0-5 years.
- 25% of loan applicants have work experience between 5-10 years.

Data Type Conversion

- Data Type for below columns were converted to ‘int’ since those were in ‘float’ :

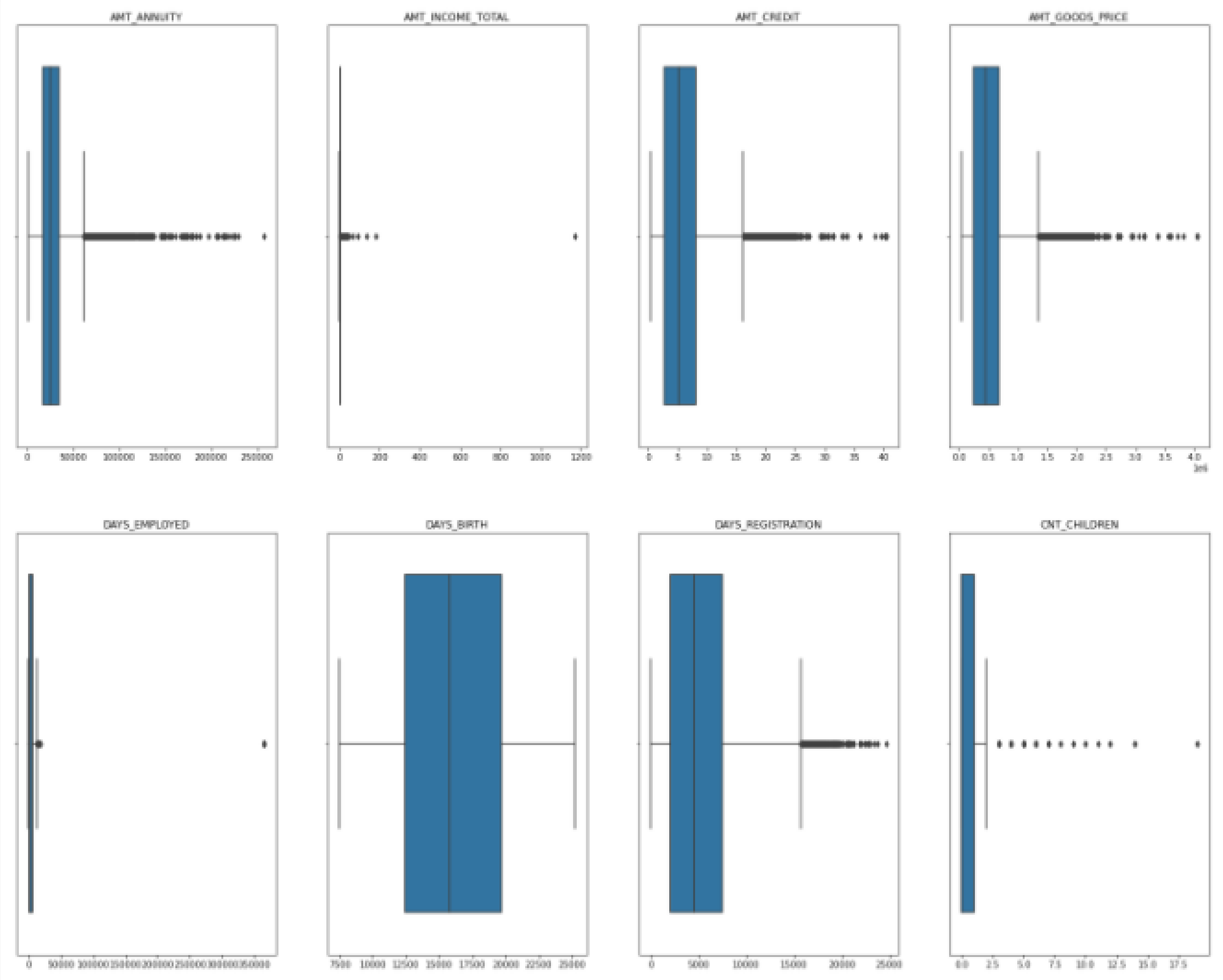
```
DAYS_BIRTH           int64
DAYS_EMPLOYED        int64
DAYS_ID_PUBLISH      int64
AGE                  int64
YEARS_EMPLOYED       int64
```

Null Value Imputation

- There were null values present in below columns and thus imputation was required:

OCCUPATION_TYPE	31.35
EMPLOYMENT_YEAR	27.08
AMT_REQ_CREDIT_BUREAU_YEAR	13.50
AMT_REQ_CREDIT_BUREAU_QRT	13.50
AMT_REQ_CREDIT_BUREAU_MON	13.50
AMT_REQ_CREDIT_BUREAU_WEEK	13.50
AMT_REQ_CREDIT_BUREAU_DAY	13.50
AMT_REQ_CREDIT_BUREAU_HOUR	13.50
NAME_TYPE_SUITE	0.42
DEF_60_CNT_SOCIAL_CIRCLE	0.33
OBS_30_CNT_SOCIAL_CIRCLE	0.33
OBS_60_CNT_SOCIAL_CIRCLE	0.33
DEF_30_CNT_SOCIAL_CIRCLE	0.33
AMT_GOODS_PRICE	0.09
AMT_INCOME_RANGE	0.08

Outlier Handling



Inference:

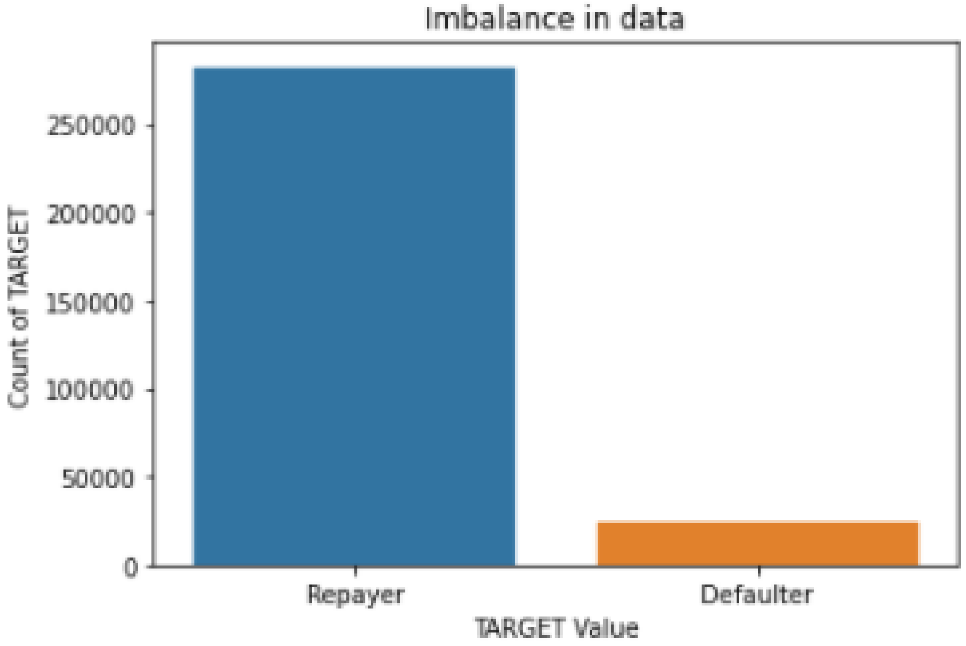
• AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have outliers.

• AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have high income.

• DAYS_BIRTH has no outliers. DAYS_EMPLOYED has outlier values around 350000(days) which is (350000//365) around 958 years which is practically impossible.

Is Data Imbalanced?

- There is data imbalance as shown below:

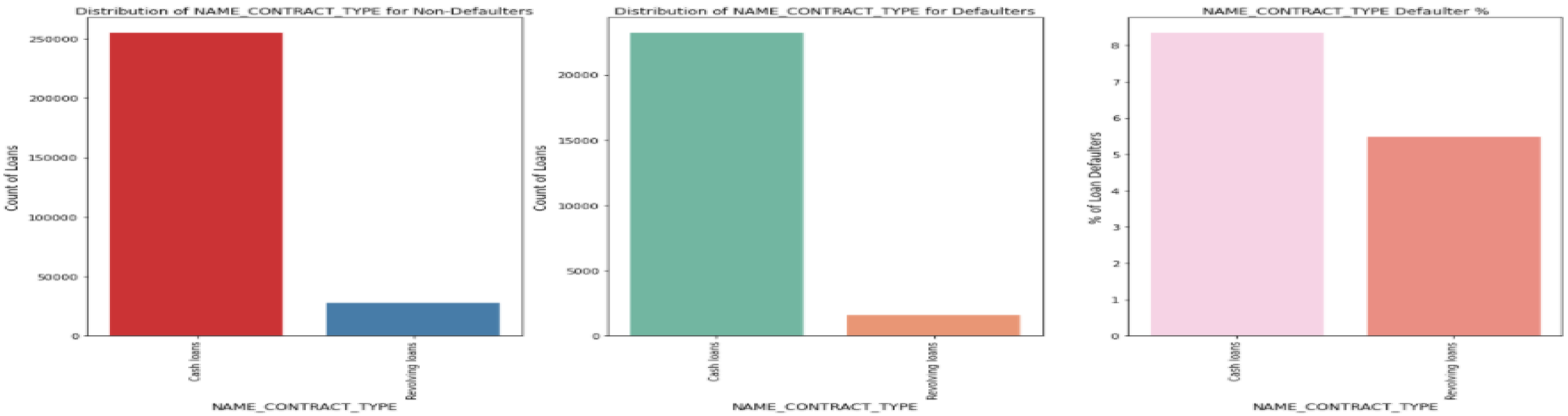


0	91.927
1	8.073

0 : Non-Defaulters / Repayers
1 : Defaulters

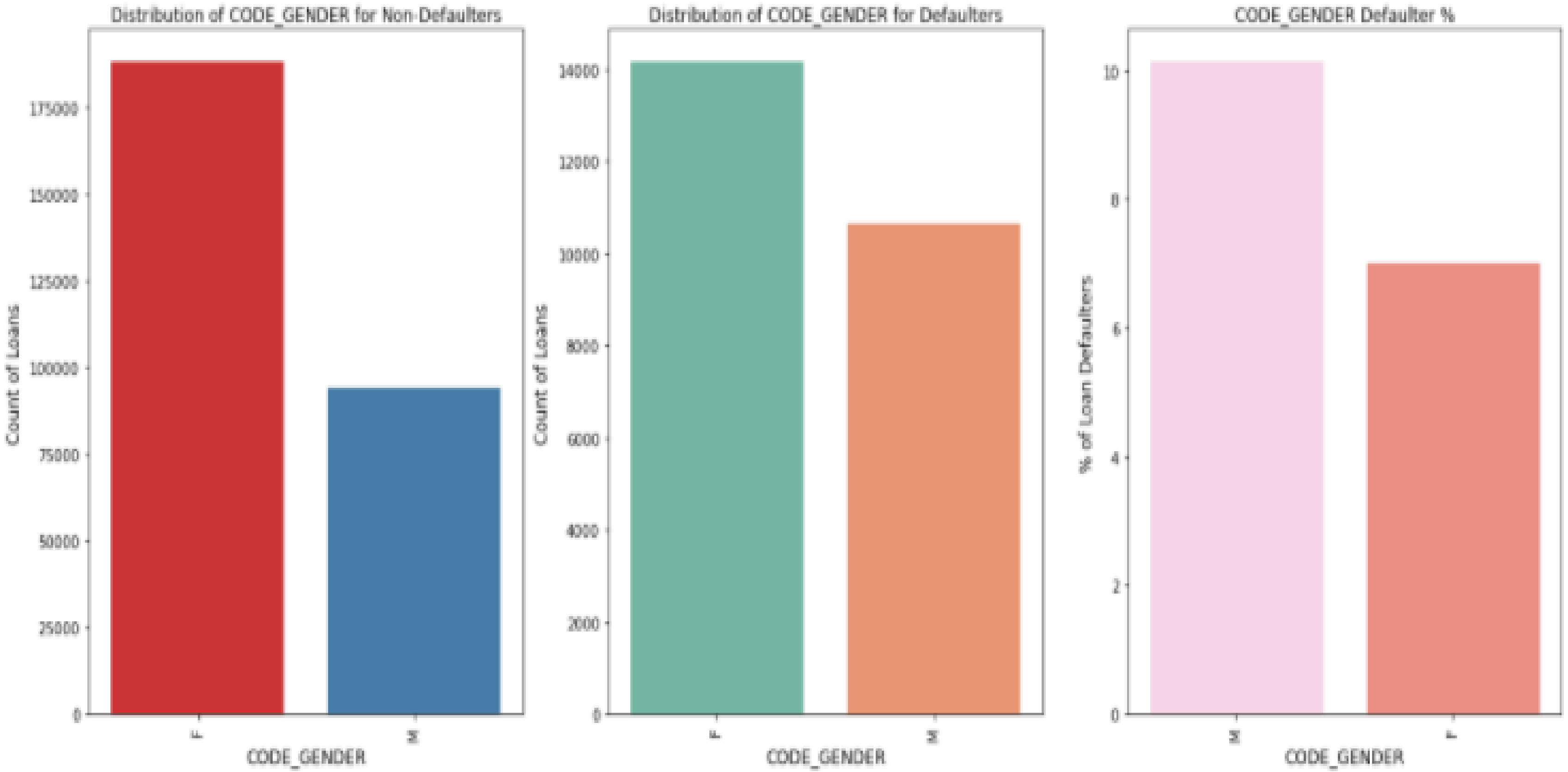
The Ratio of Data Imbalance is - **11.39 : 1**

Univariate Analysis – 'NAME_CONTRACT_TYPE' column



Inference
Cash Loans are higher in number than Revolving Loans for both Defaulters and Non-Defaulters.
Cash Loans contract type has the maximum percentage of Loan Payment Difficulties

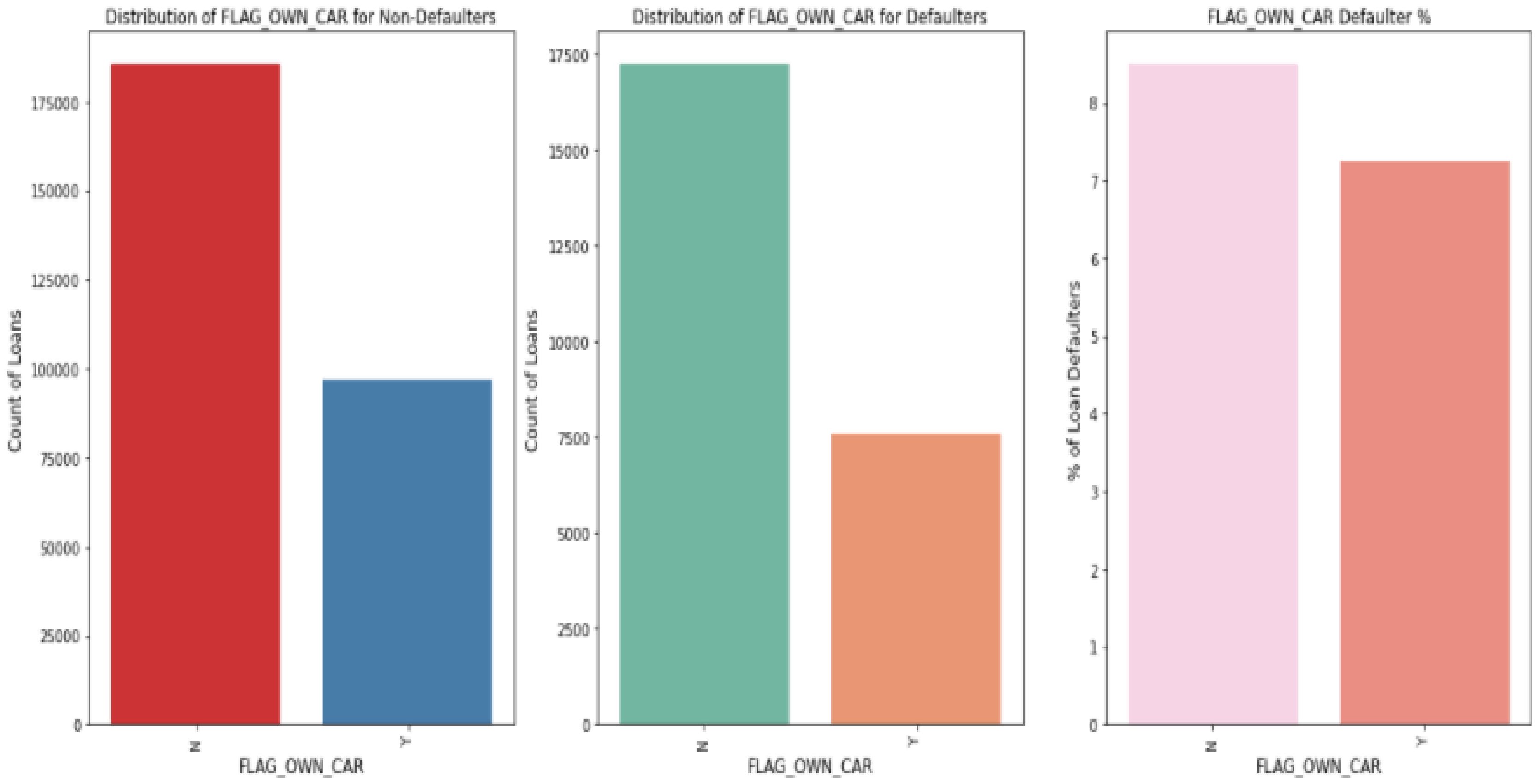
Univariate Analysis – ‘CODE_GENDER’ column



Inference

- Number of Females taking Loans is much higher than the Number of Males for both Defaulters and Non-Defaulters.
- Males have a higher chance of defaulting than Females.

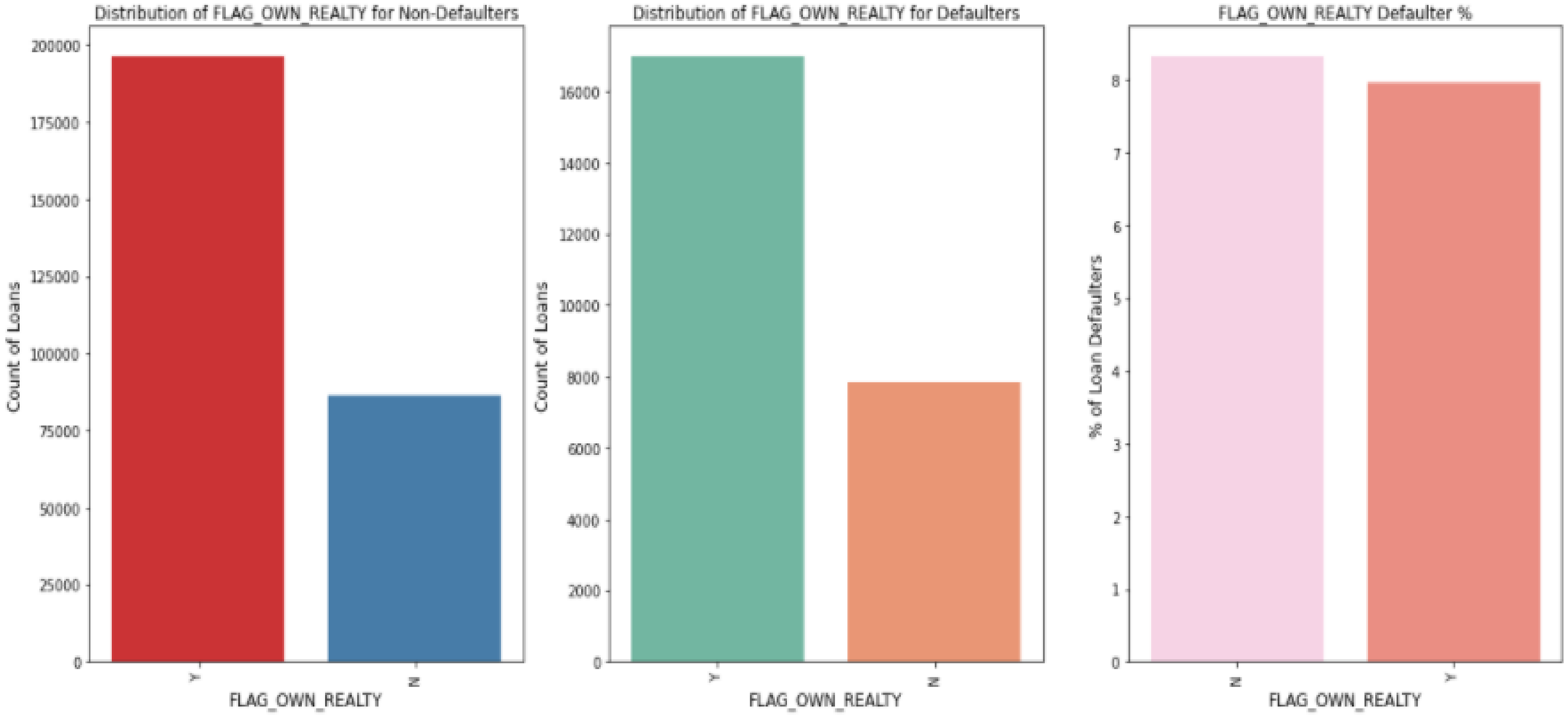
Univariate Analysis – 'FLAG_OWN_CAR' column



Inference

- Most number of people applying for Loan don't own a car.
- People not owning a car have a slightly higher default rate than people who owns a car, though there is not much correlation.

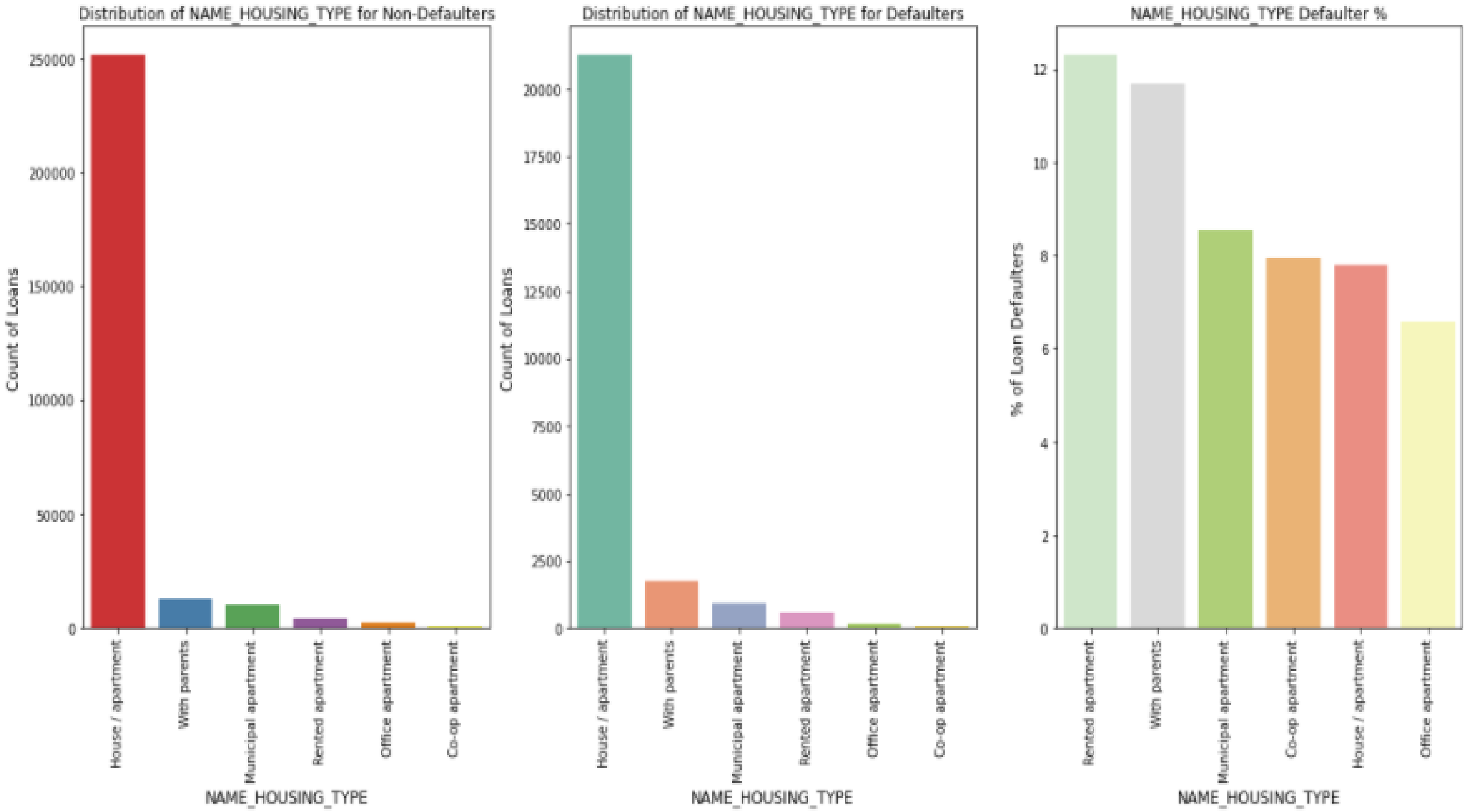
Univariate Analysis – 'FLAG_OWN_REALTY' column



Inference

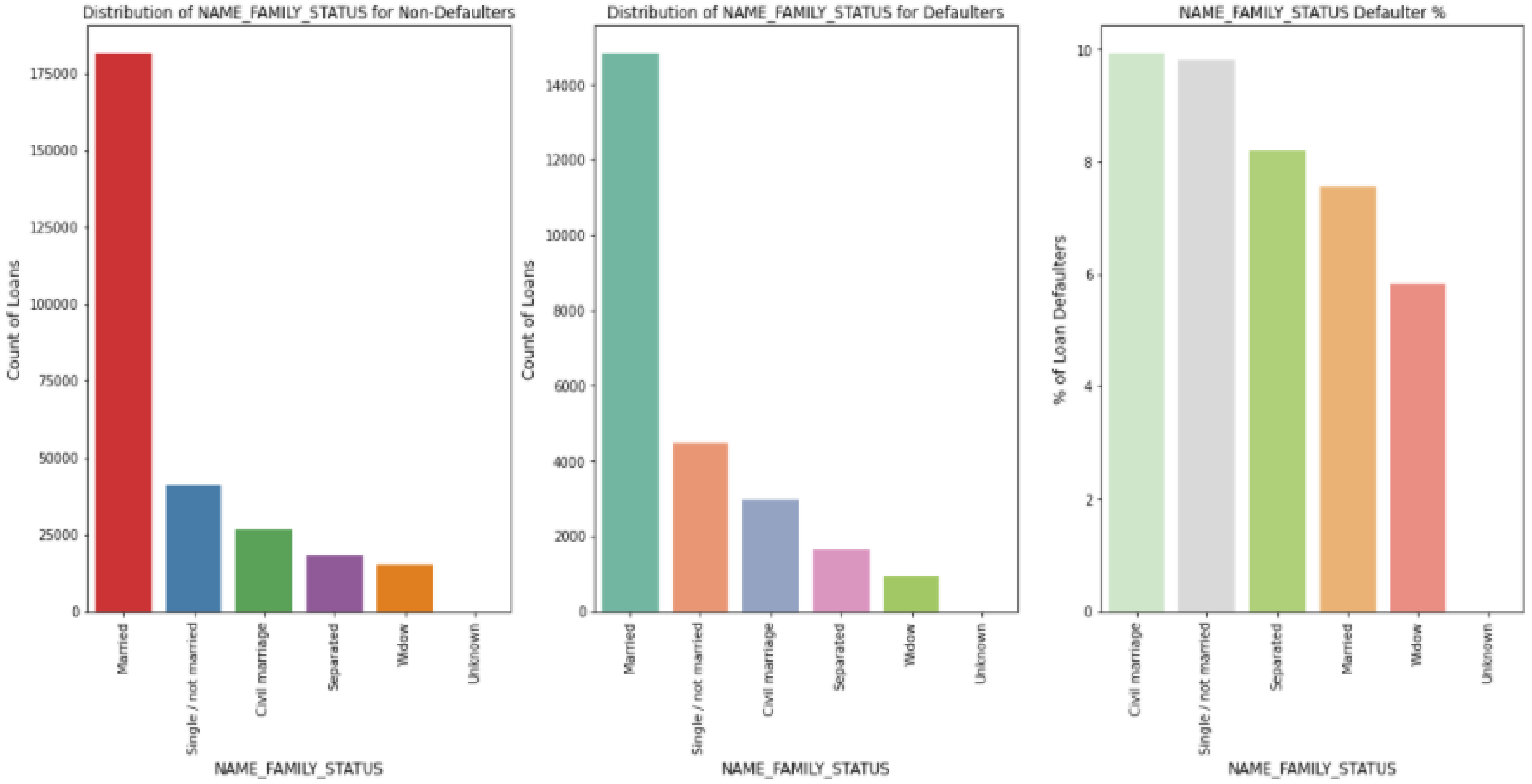
- Most number of people applying for Loan owns a house or flat.
- Defaulting rate of both categories are more or less same, i.e. there is no correlation between owning a house and defaulting a loan.

Univariate Analysis – ‘NAME_HOUSING_TYPE’ column



Inference:
- Most of the people live in a house/apartment.
- People living in rented apartments and people living with their parents have higher probabilities of defaulting.
- People living in office apartments have the lowest defaulting rate.

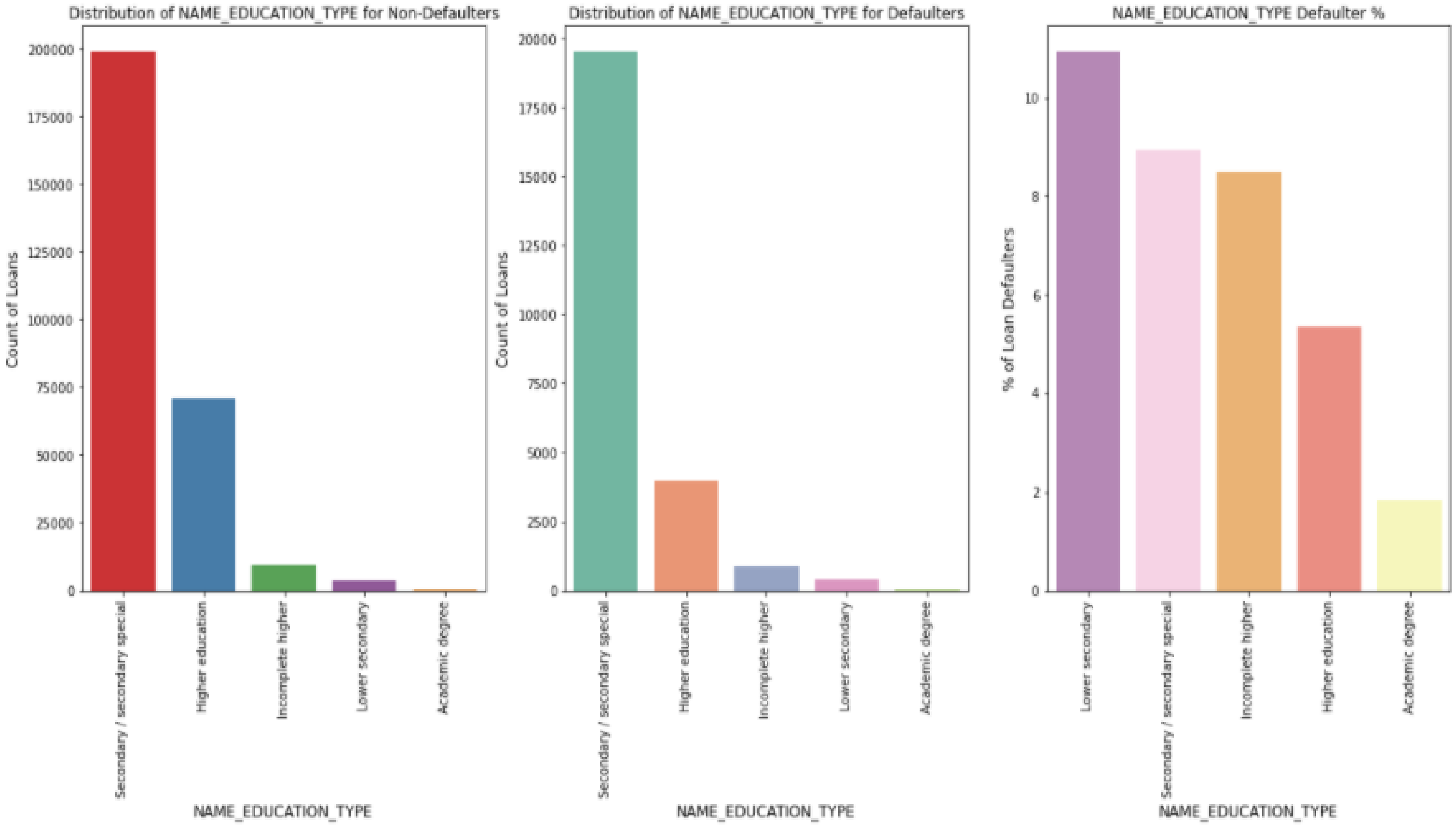
Univariate Analysis – ‘NAME_FAMILY_STATUS’ column



Inference:

- Most people who have taken loan are married.
- Civil marriage has the highest rate of defaulting followed by single/not married people.
- Widow has the lowest rate of defaulting.

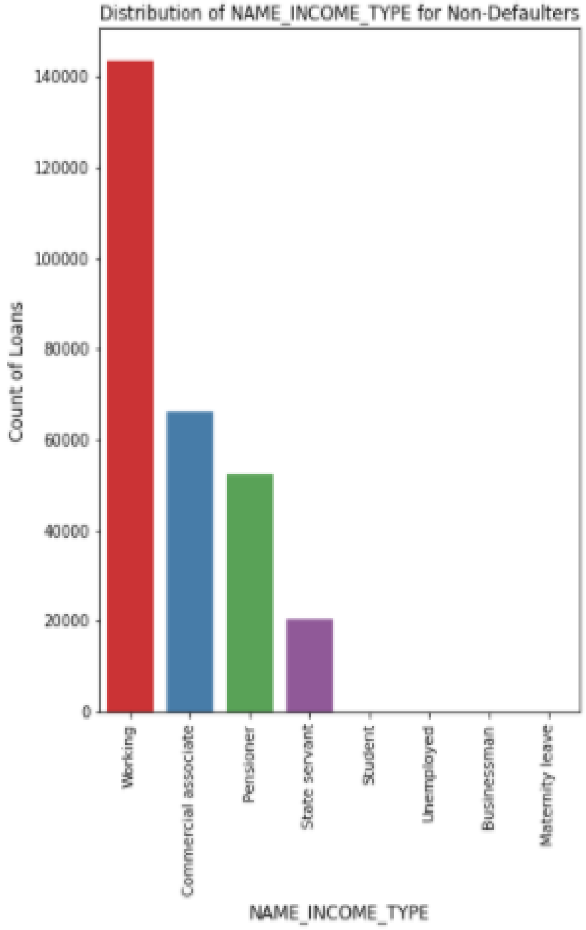
Univariate Analysis – 'NAME_EDUCATION_TYPE' column

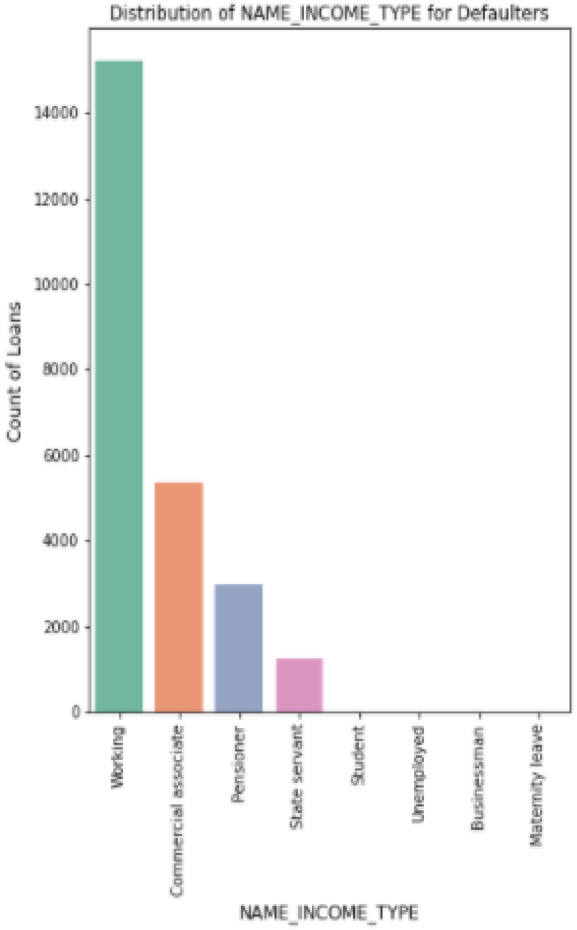


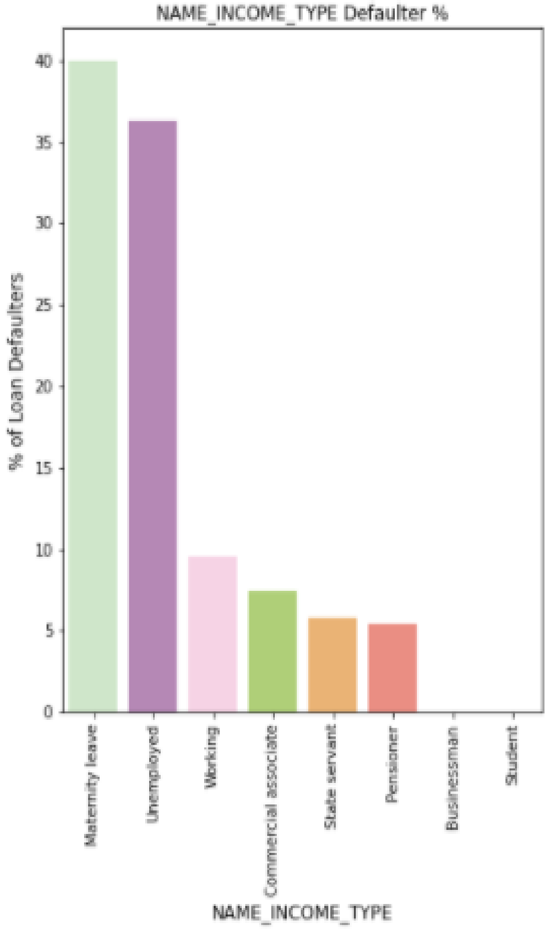
Inference:

- People with academic degree rarely take loans. Also, they are rare defaulters. So, they are potentially good customers.
- People with higher education are less likely to have payment difficulties.
- Lower secondary category has the highest rate of defaulting.

Univariate Analysis – ‘NAME_INCOME_TYPE’ column



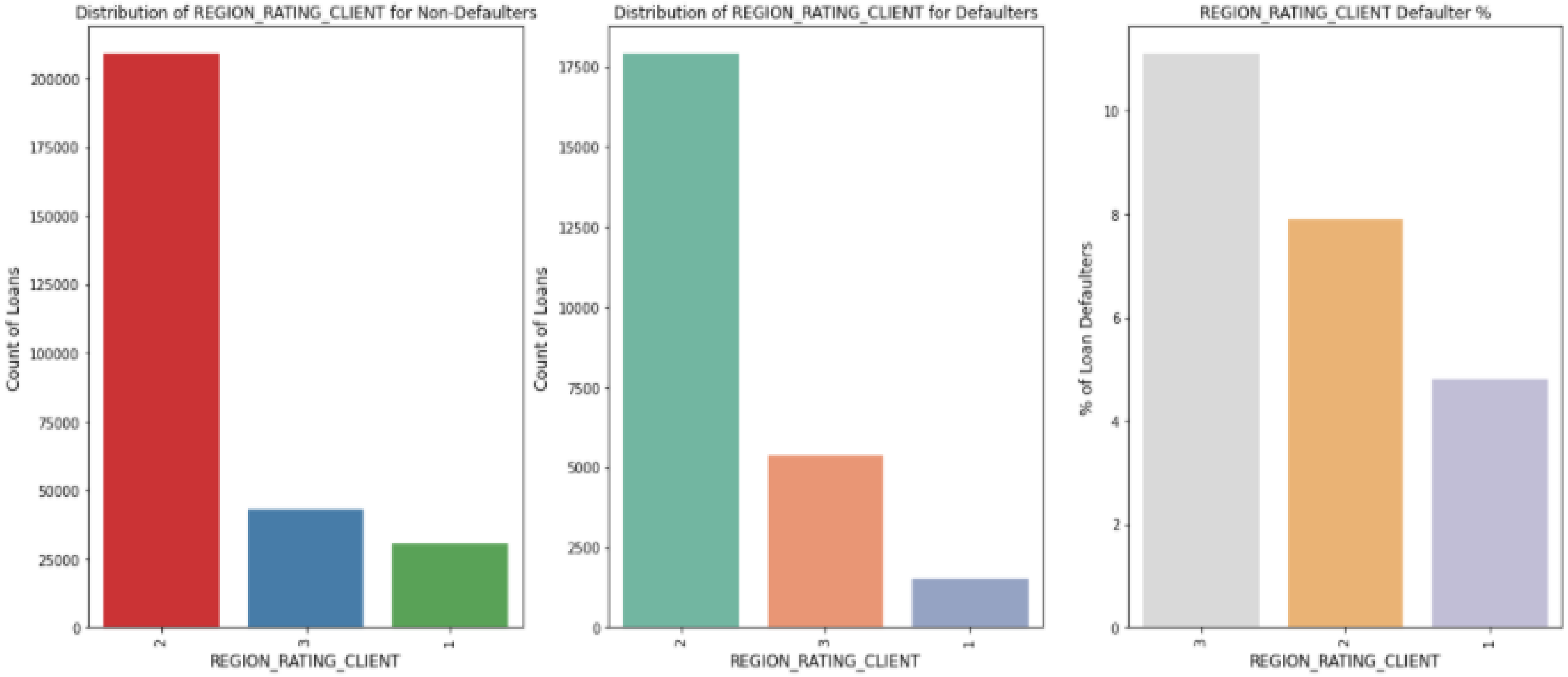




Inference:

- Most of the applicants for loans have income type as 'Working'.
- Maternity leave income type has the highest rate of defaulting, followed by Unemployed people.
- Student and Businessman are good categories to target since they don’t have any default record.

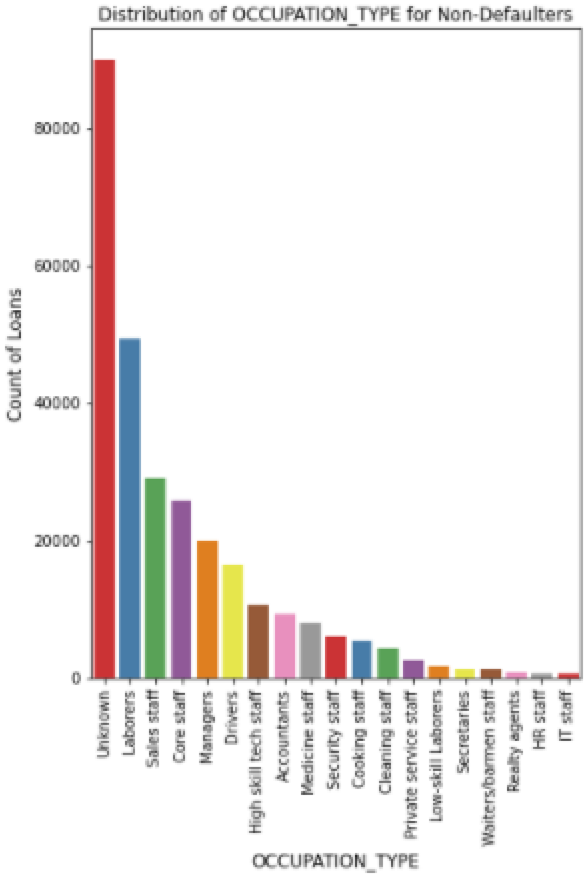
Univariate Analysis – ‘REGION_RATING_CLIENT’ column

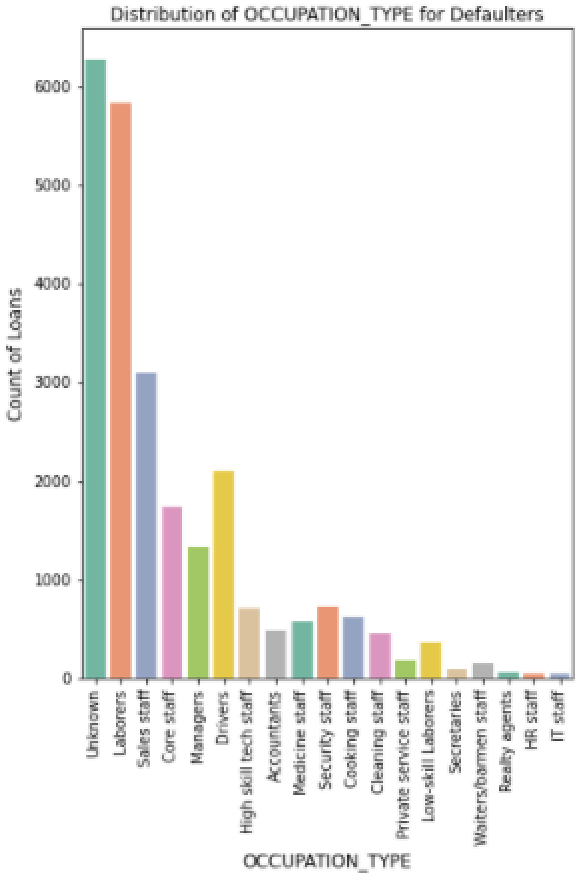


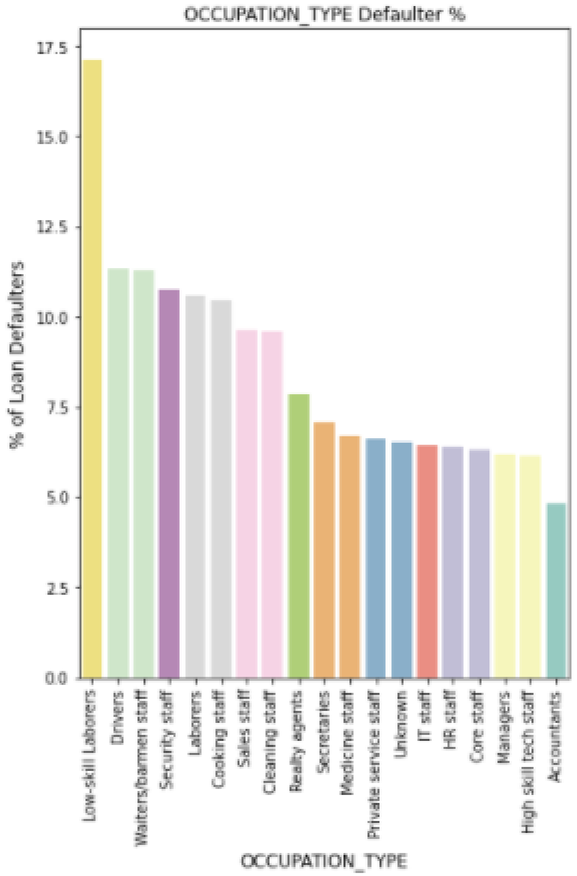
Inference:

- Region 3 has the highest default rate.
- Region 1 has the lowest default rate.

Univariate Analysis – ‘OCCUPATION_TYPE’ column

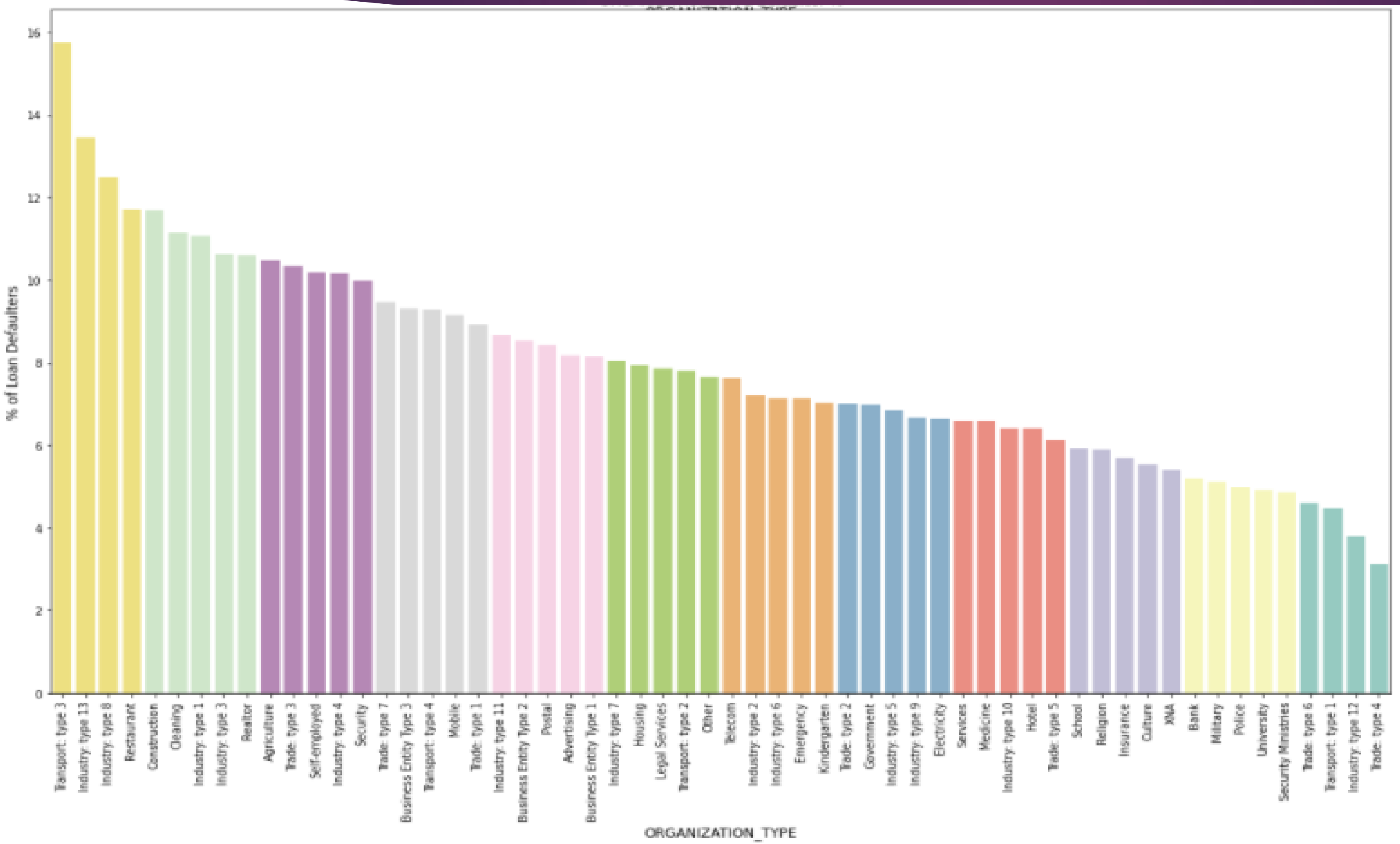






Inference:
- Most of the loans are taken by Laborers, followed by Sales staff. (Excluding 'Unknown')
- Low skill laborers has the highest default rate, followed by Drives, Waiters, Security staff.

Univariate Analysis – ‘ORGANIZATION_TYPE’ column



Inference:

• Most of the loans are taken by Business Entity Type 3.

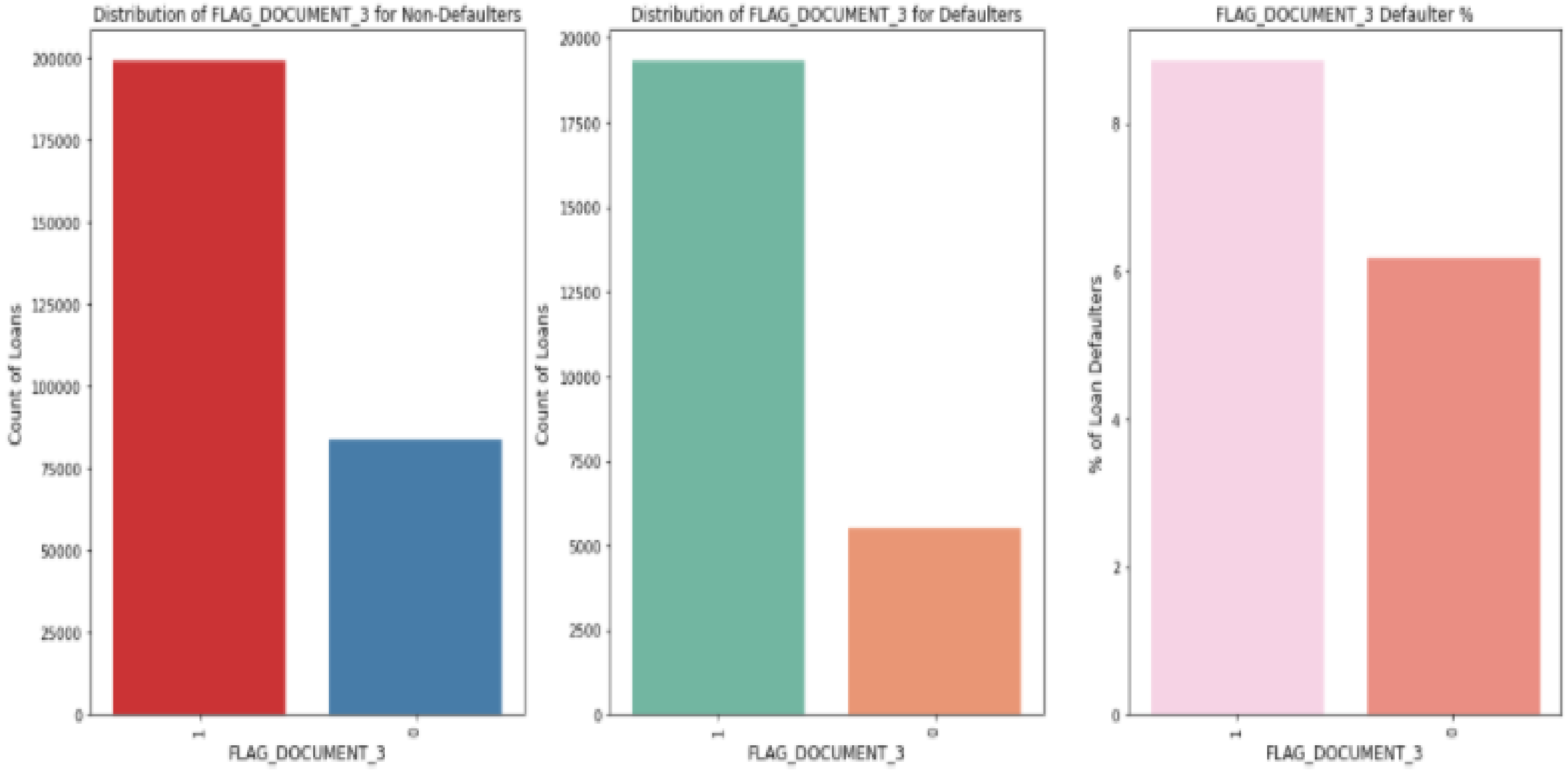
• Organization type information is unavailable for many of the loan applicants.

• Transport Type 3 has around 16% default rate.

• Industry Type 13 has around 13.8% default rate, followed by Industry Type 8, Restaurant, Construction.

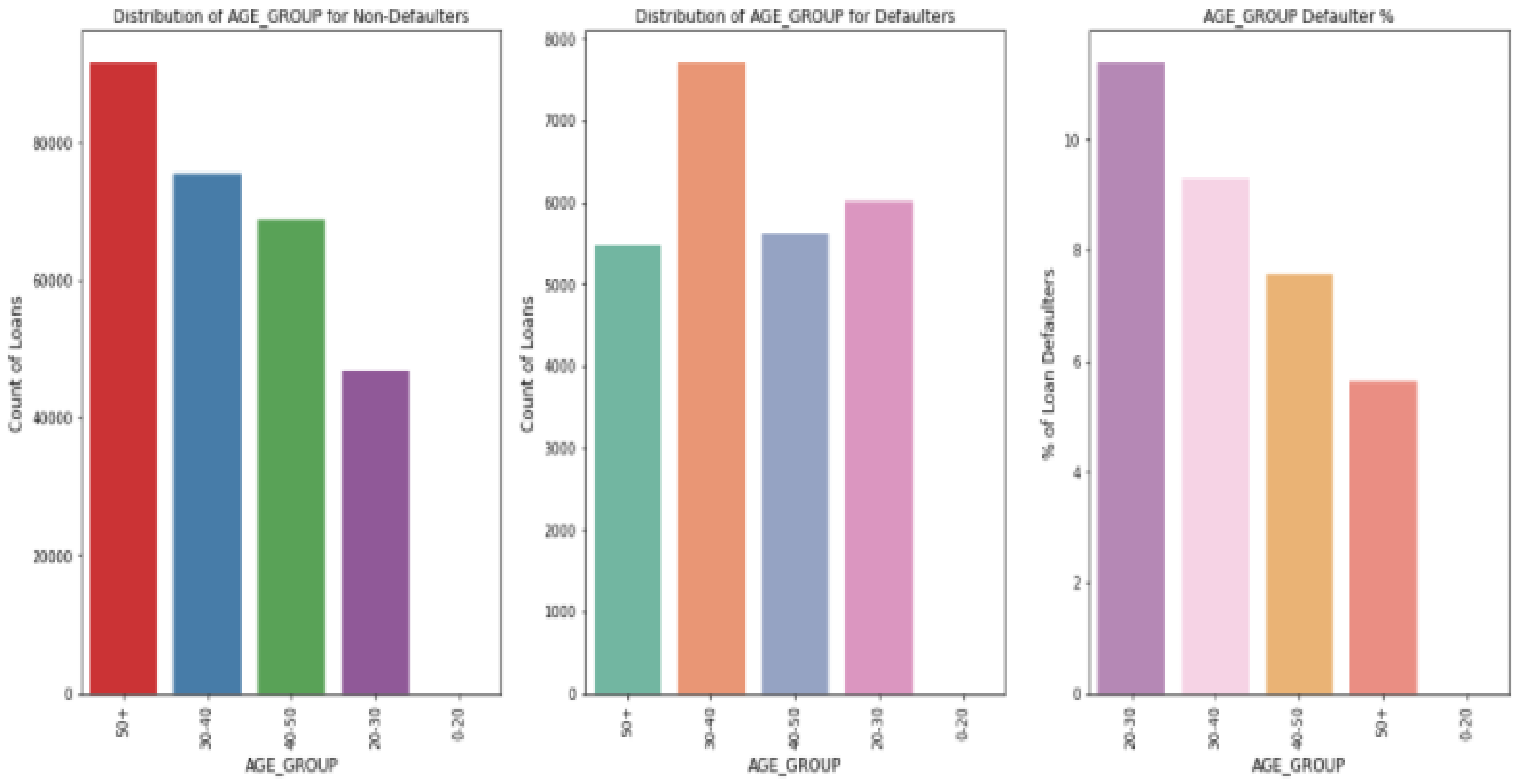
• Trade Type 4 has the lowest default rate (3%).

Univariate Analysis – ‘FLAG_DOCUMENT_3’ column



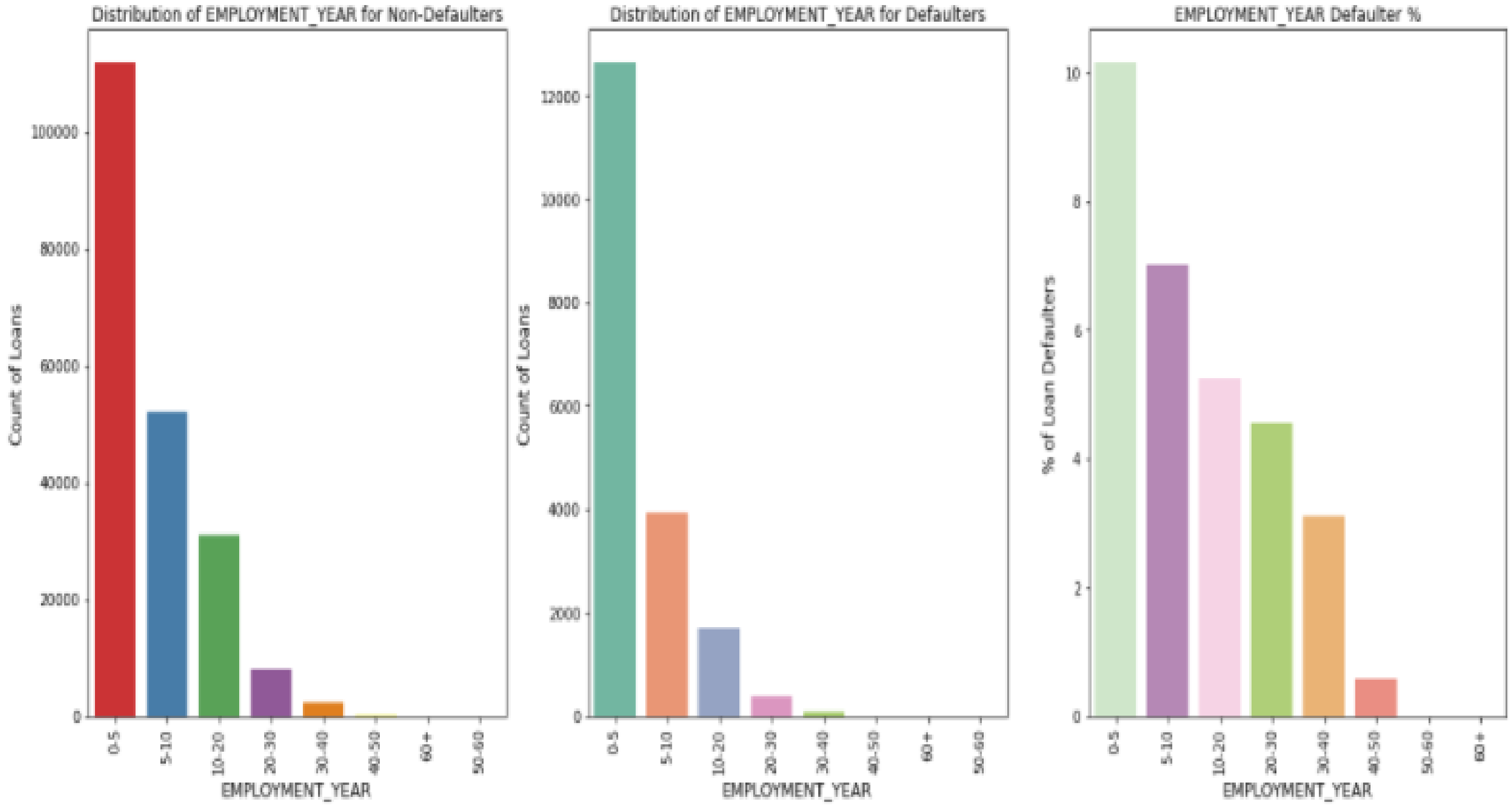
Inference:
No correlation, even if applicant submitted the document they have defaulted slightly more than the ones who haven't submitted.

Univariate Analysis – 'AGE_GROUP' column



Inference:
• 20-30 Age group people have a higher default rate.
• 50+ Age group have a lower default rate.

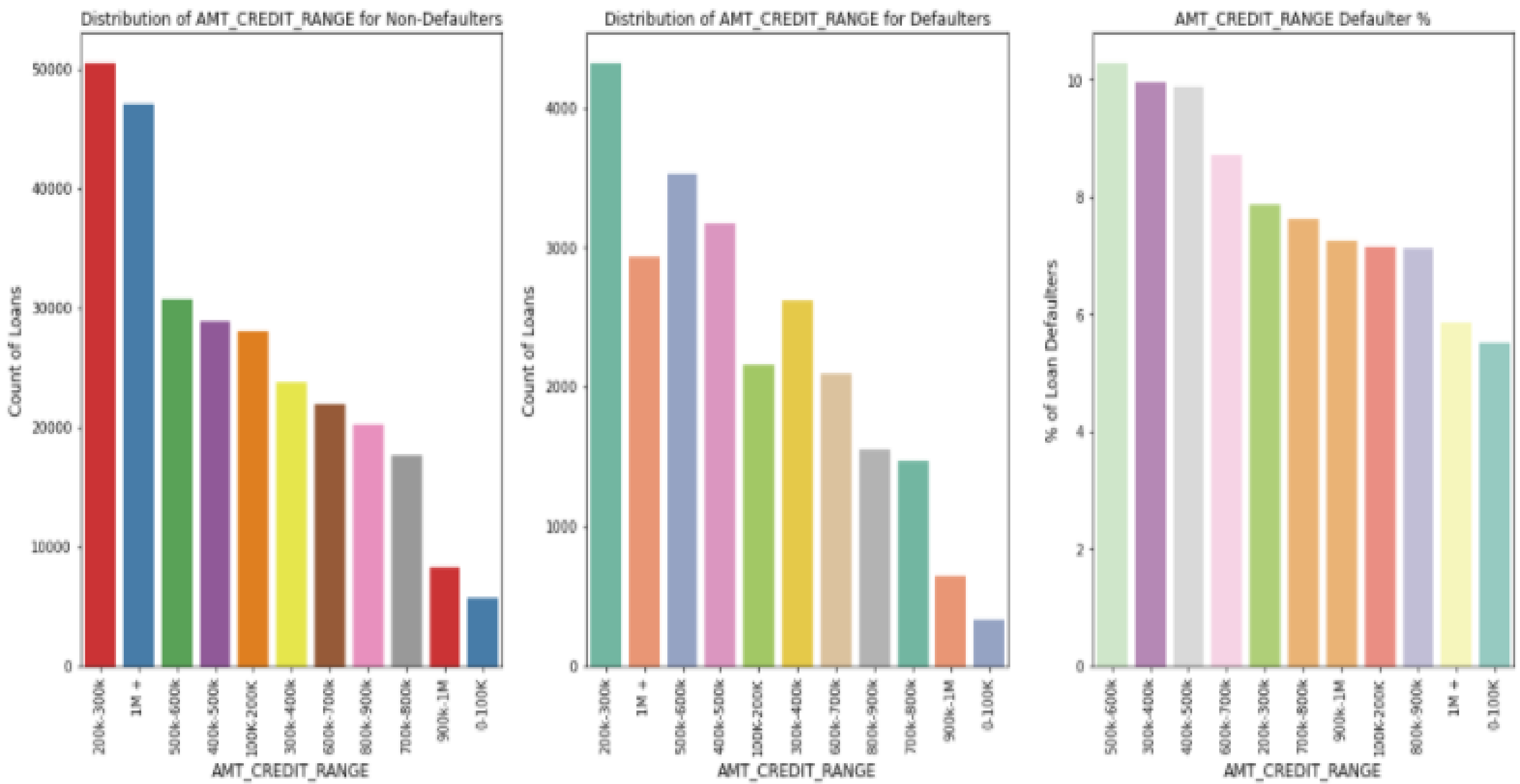
Univariate Analysis – 'EMPLOYMENT_YEAR' column



Inference:

- Majority of the applicants have been employed in 0-5 years.
- With increase in employment year, defaulting rate is also decreasing.

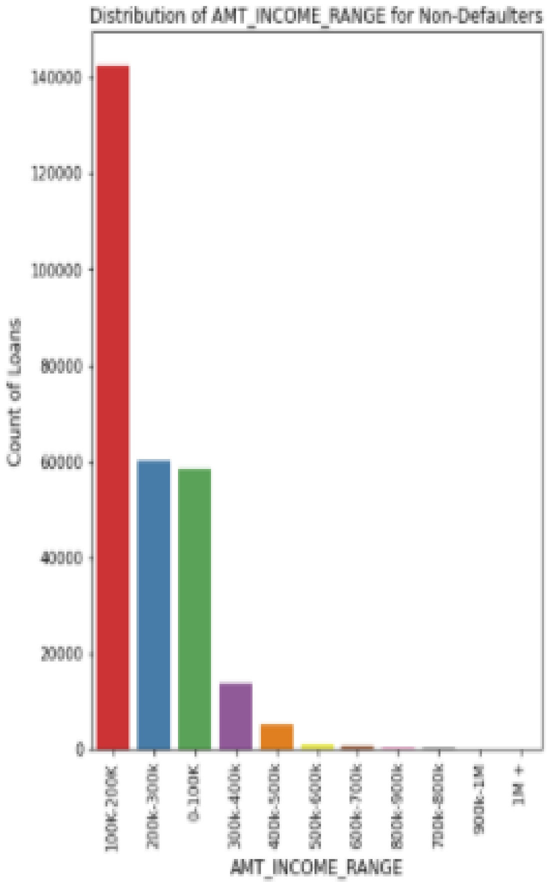
Univariate Analysis – ‘AMT_CREDIT_RANGE’ column

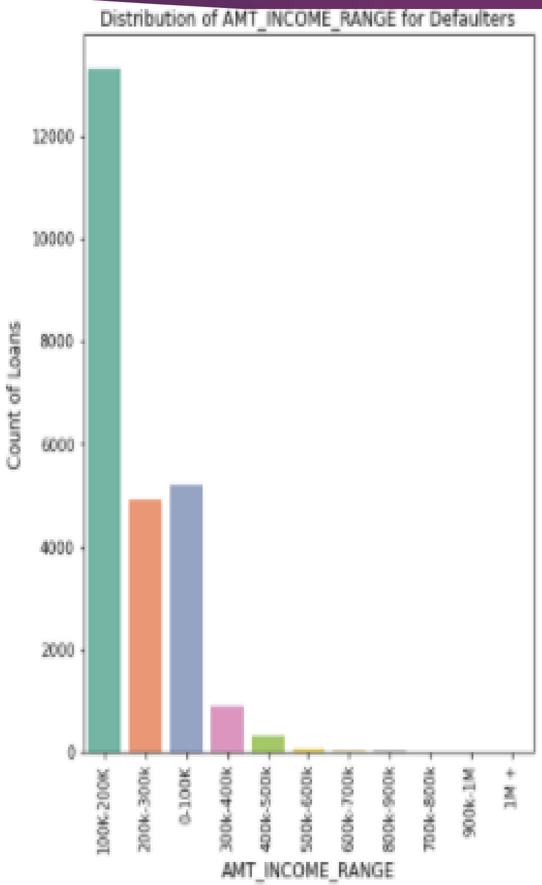


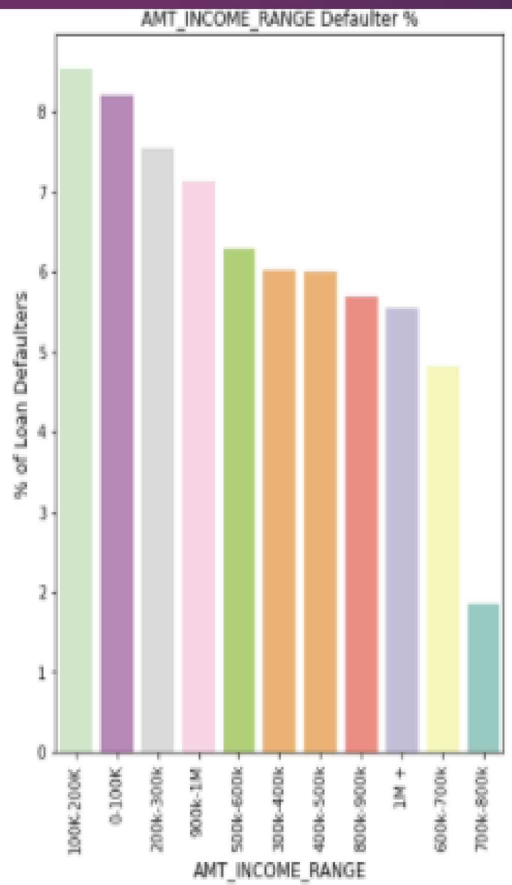
Inference:

- People who get loan of 0-100K defaults less, followed by people who receive a loan of more than 1M.
- People who get loan of 500k-600k have a higher default rate.

Univariate Analysis – ‘AMT_INCOME_RANGE’ column



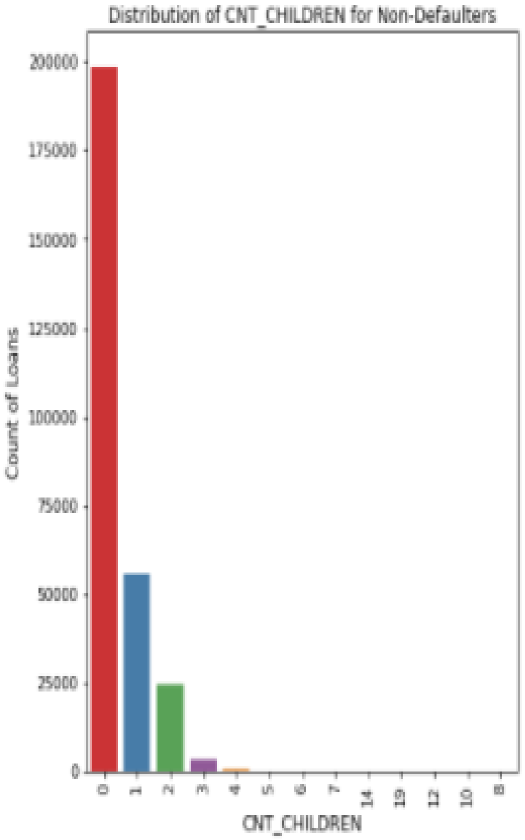


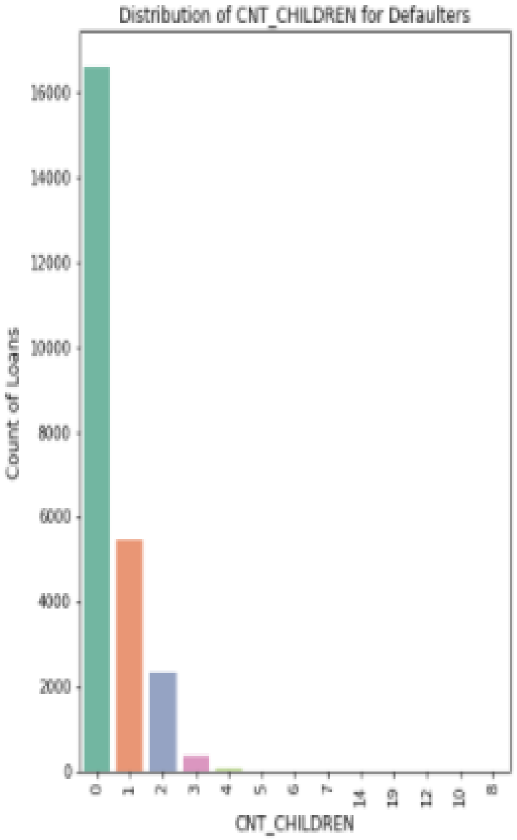


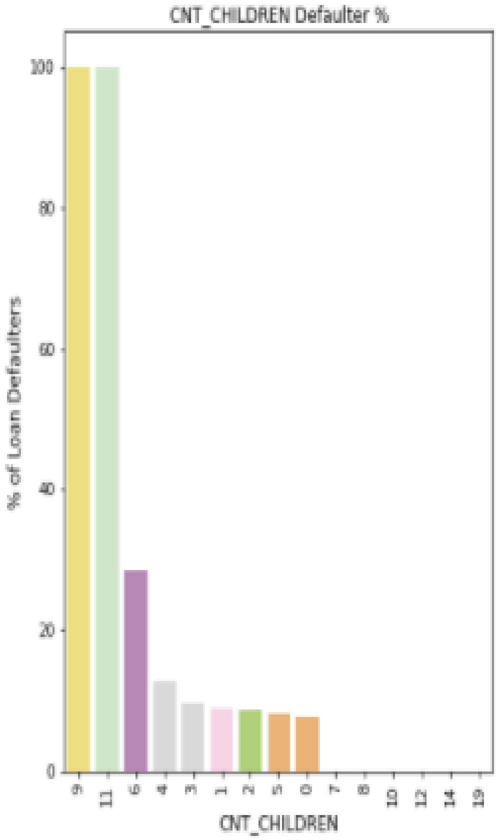
Inference:

- People with income less than 300k has a higher probability of defaulting.
- People with income more than 700k are less likely to default.

Univariate Analysis – ‘CNT_CHILDREN’ column







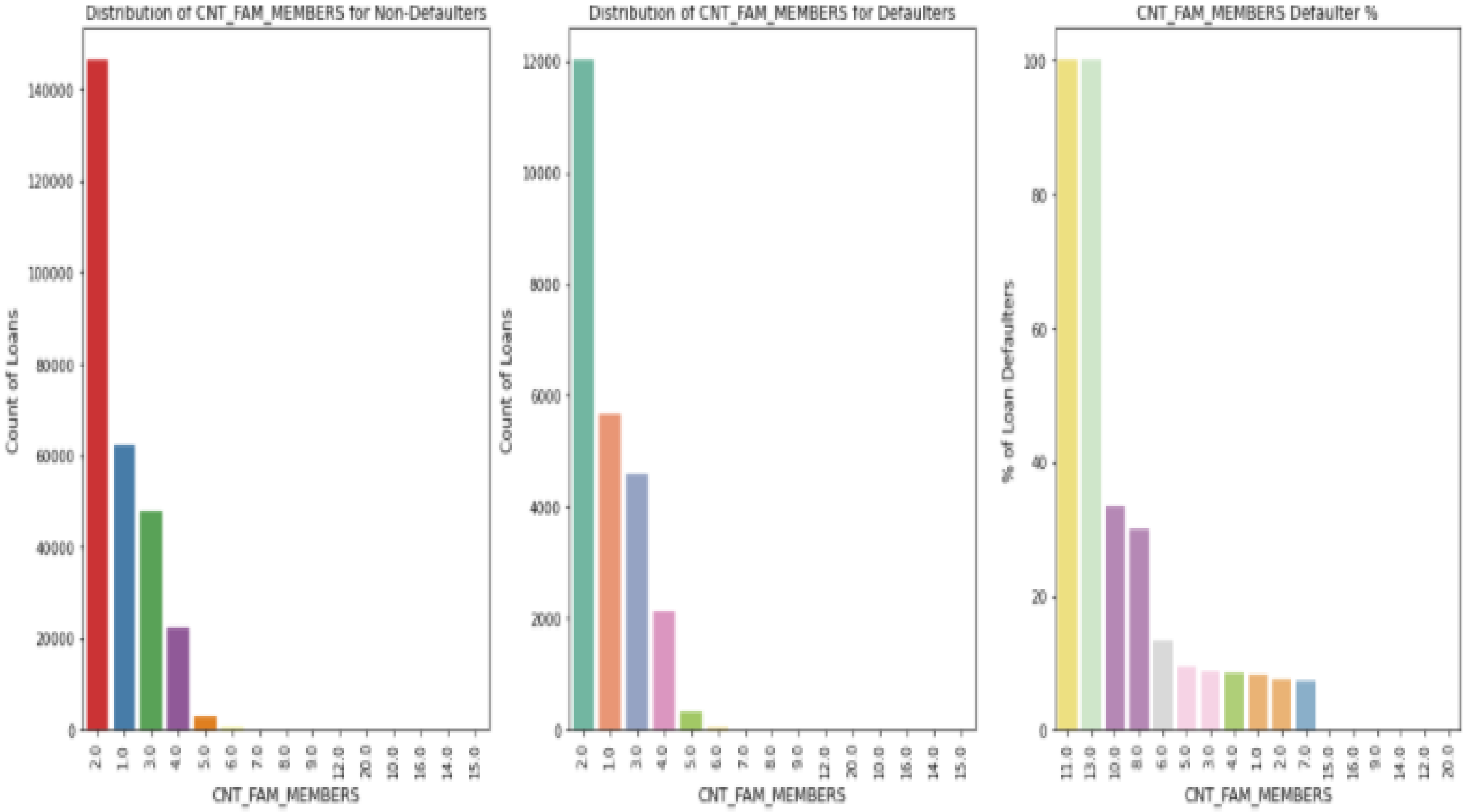
Inference:

• Most of the applicants don’t have any children.

• People with more than 4 children have a very high probability of defaulting.

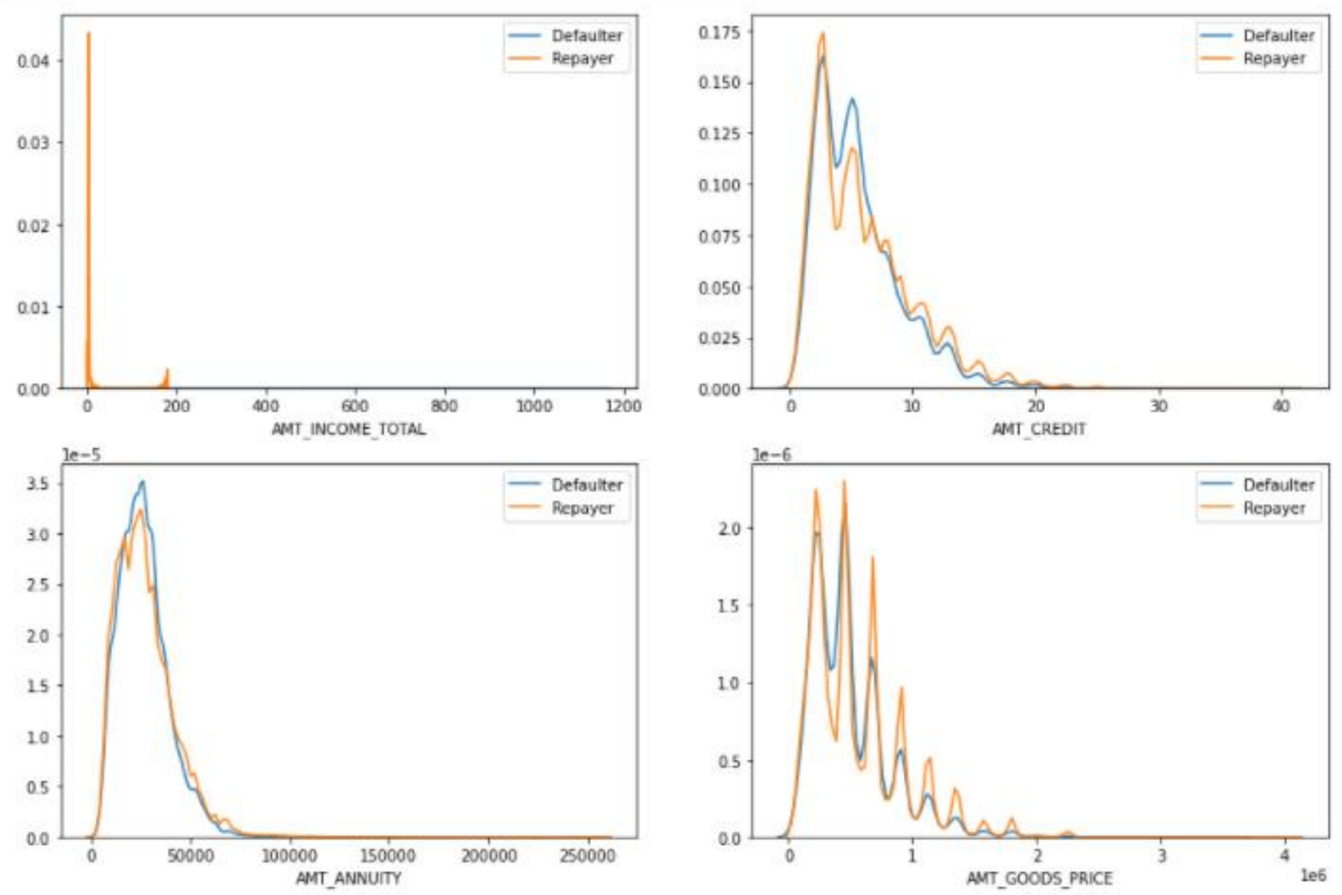
People with 9 and 11 children have 100% default rate.

Univariate Analysis – ‘CNT_FAM_MEMBERS’ column



Inference:
Having more family members increases the risk of defaulting.

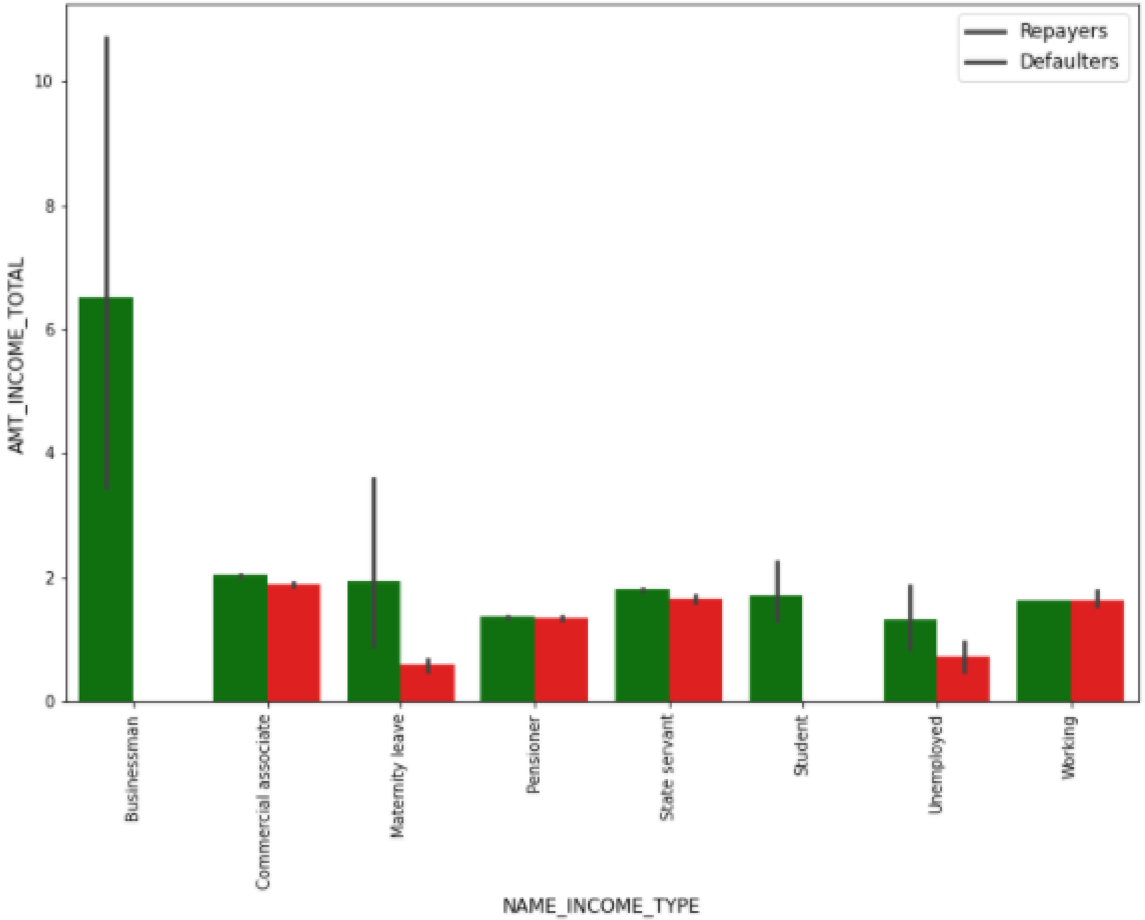
Univariate Analysis - Numerical



Inference:

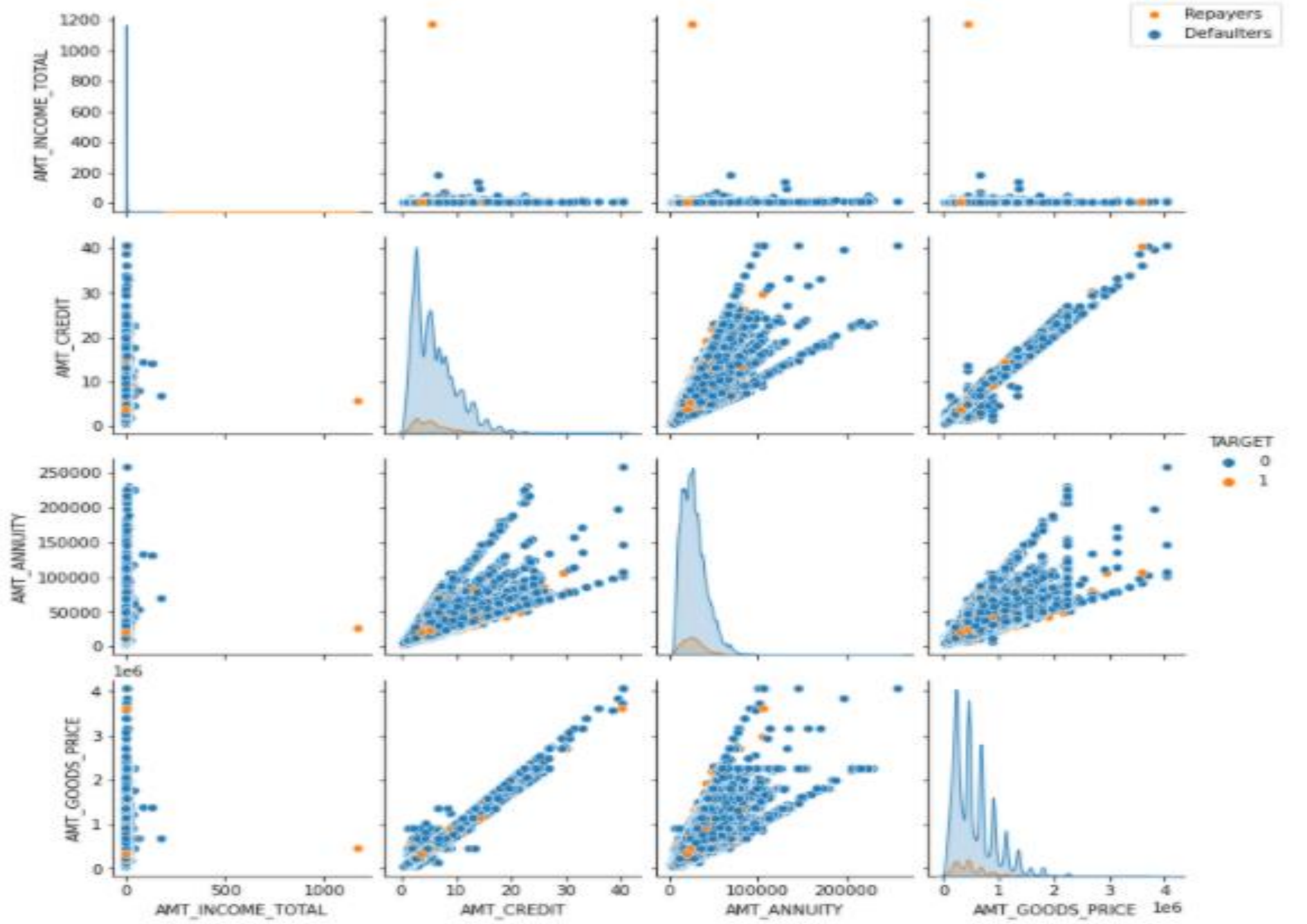
- Most people pay annuity below 50k for the credit loan.
- Credit amount of the loan is mostly less then 10 lakhs.
- The repayers and defaulters distribution overlap in all the plots. We cannot use any of these variables to make a decision.

Bivariate Analysis - NAME_INCOME_TYPE vs. AMT_INCOME_TOTAL



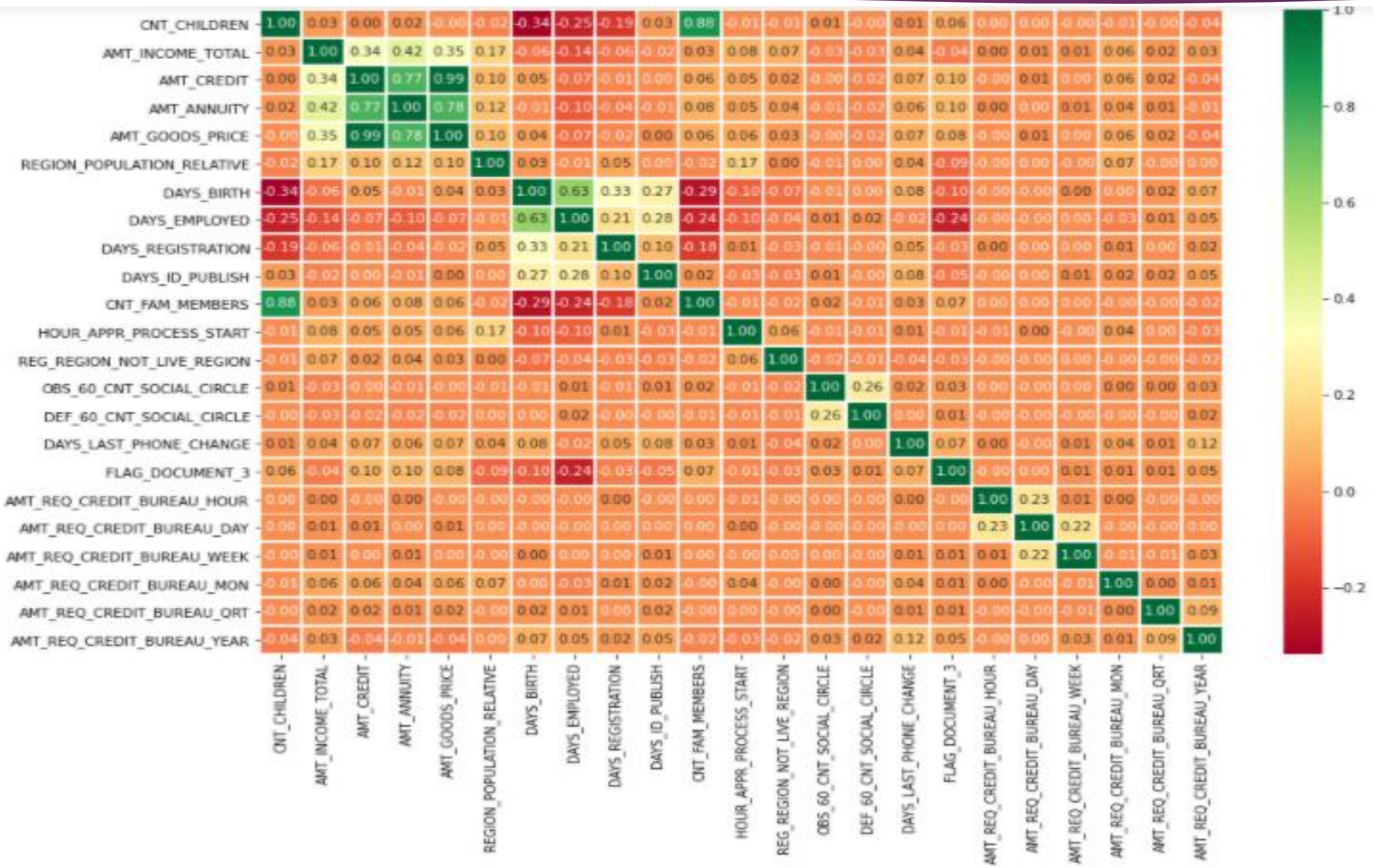
Inference:
Businessman's income is the highest.

Bivariate Analysis - Numerical



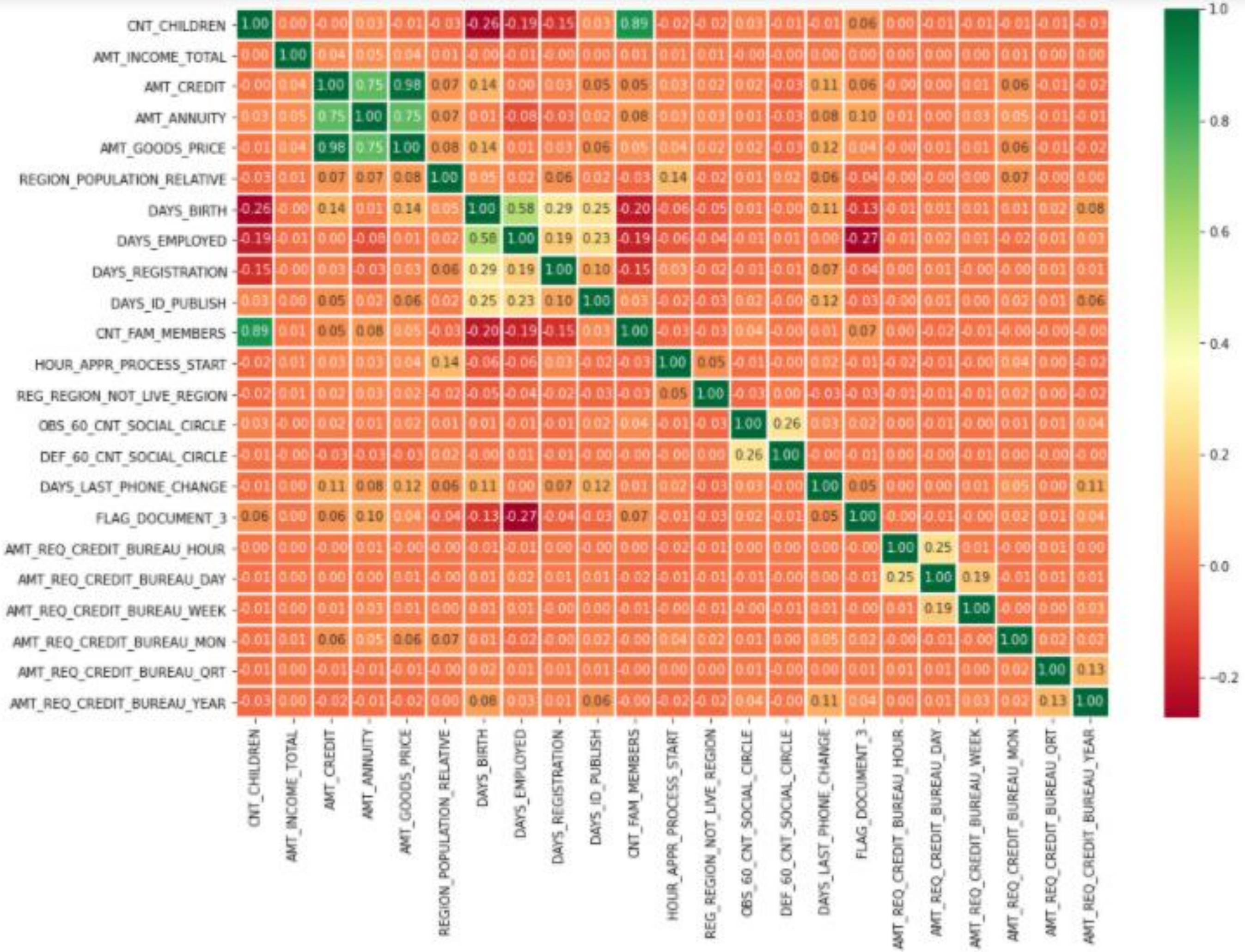
Inference:
- 'AMT_CREDIT' and 'AMT_GOODS_PRICE' are highly correlated.
- Very less defaulters for 'AMT_CREDIT' > 3M.
- When 'AMT_ANNUITY' > 150K chances of defaulting is low.

Correlation b/w Numeric Variables for Repayers



Inference:
Credit amount is highly correlated with 'Amount of Goods Price', 'Amount Annuity', 'Total Income'. Repayers have high correlation with 'Number of Days Employed'.

Correlation b/w Numeric Variables for Defaulters



Inference:

- Credit amount is highly correlated with 'Amount of Goods Price' which is same as Repayers.
- Loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to Repayers(0.77).
- Repayers have high correlation in number of days employed(0.62) when compared to defaulters(0.58).
- Drop in the correlation between total income of the client and the credit amount(0.038) amongst defaulters whereas it is 0.342 among repayers.
- Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in repayers.
- Increase in defaulted to observed count in social circle among defaulters(0.264) when compared to repayers(0.254)

Previous Data Analysis

Dimension of Previous Dataframe: (1670214, 37)

Structure of Data frame

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1670214 entries, 0 to 1670213
Data columns (total 37 columns):
 #   Column                       Non-Null Count    Dtype
---  ------                       --------------    -----
 0   SK_ID_PREV                   1670214 non-null  int64
 1   SK_ID_CURR                   1670214 non-null  int64
 2   NAME_CONTRACT_TYPE           1670214 non-null  object
 3   AMT_ANNUITY                  1297979 non-null  float64
 4   AMT_APPLICATION              1670214 non-null  float64
 5   AMT_CREDIT                   1670213 non-null  float64
 6   AMT_DOWN_PAYMENT             774370 non-null   float64
 7   AMT_GOODS_PRICE              1284699 non-null  float64
 8   WEEKDAY_APPR_PROCESS_START   1670214 non-null  object
 9   HOUR_APPR_PROCESS_START      1670214 non-null  int64
 10  FLAG_LAST_APPL_PER_CONTRACT  1670214 non-null  object
 11  NFLAG_LAST_APPL_IN_DAY       1670214 non-null  int64
 12  RATE_DOWN_PAYMENT            774370 non-null   float64
 13  RATE_INTEREST_PRIMARY        5951 non-null     float64
 14  RATE_INTEREST_PRIVILEGED     5951 non-null     float64
 15  NAME_CASH_LOAN_PURPOSE       1670214 non-null  object
 16  NAME_CONTRACT_STATUS         1670214 non-null  object
 17  DAYS_DECISION                1670214 non-null  int64
 18  NAME_PAYMENT_TYPE            1670214 non-null  object
 19  CODE_REJECT_REASON           1670214 non-null  object
 20  NAME_TYPE_SUITE              849809 non-null   object
 21  NAME_CLIENT_TYPE             1670214 non-null  object
 22  NAME_GOODS_CATEGORY          1670214 non-null  object
 23  NAME_PORTFOLIO               1670214 non-null  object
 24  NAME_PRODUCT_TYPE            1670214 non-null  object
 25  CHANNEL_TYPE                 1670214 non-null  object
 26  SELLERPLACE_AREA             1670214 non-null  int64
 27  NAME_SELLER_INDUSTRY         1670214 non-null  object
 28  CNT_PAYMENT                  1297984 non-null  float64
 29  NAME_YIELD_GROUP             1670214 non-null  object
 30  PRODUCT_COMBINATION          1669868 non-null  object
 31  DAYS_FIRST_DRAWING           997149 non-null   float64
 32  DAYS_FIRST_DUE               997149 non-null   float64
 33  DAYS_LAST_DUE_1ST_VERSION    997149 non-null   float64
 34  DAYS_LAST_DUE                997149 non-null   float64
 35  DAYS_TERMINATION             997149 non-null   float64
 36  NFLAG_INSURED_ON_APPROVAL    997149 non-null   float64
dtypes: float64(15), int64(6), object(16)
memory usage: 471.5+ MB
```

Columns with Missing value % >= 40%

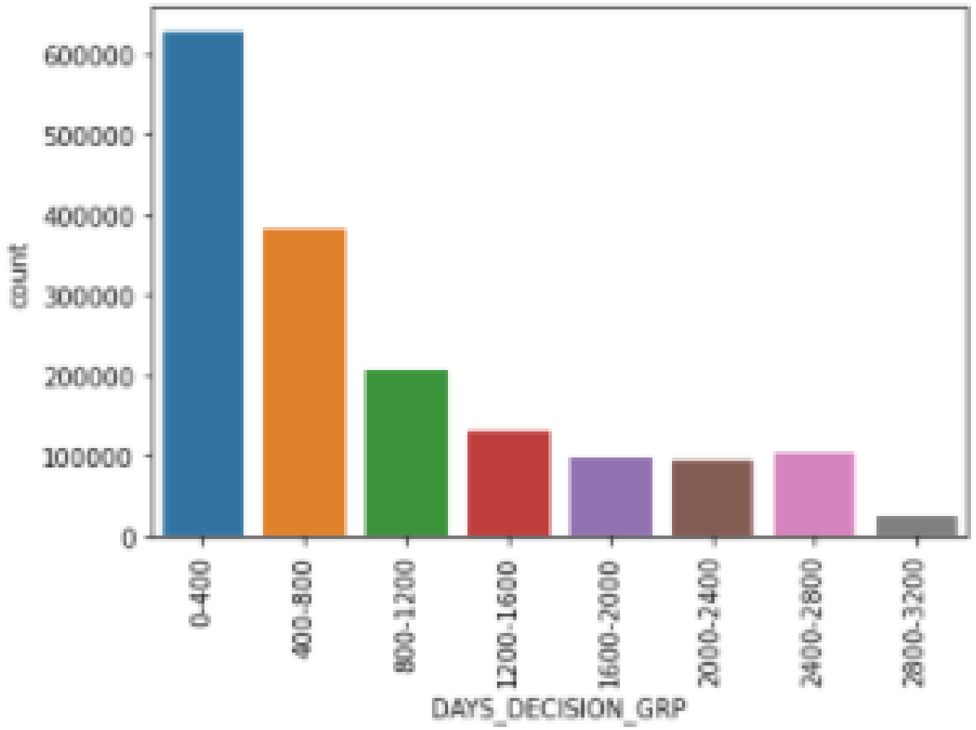
```
RATE_INTEREST_PRIVILEGED     99.64
RATE_INTEREST_PRIMARY        99.64
RATE_DOWN_PAYMENT            53.64
AMT_DOWN_PAYMENT             53.64
NAME_TYPE_SUITE              49.12
DAYS_TERMINATION             40.30
NFLAG_INSURED_ON_APPROVAL    40.30
DAYS_FIRST_DRAWING           40.30
DAYS_FIRST_DUE               40.30
DAYS_LAST_DUE_1ST_VERSION    40.30
dtype: float64
```

Insight:
There are 11 columns which have more than or equal to 40% missing values.

Post dropping this columns the new dimensions of Dataframe became: (1670214, 22)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1670214 entries, 0 to 1670213
Data columns (total 22 columns):
 #   Column                  Non-Null Count    Dtype
---  ------                  --------------    -----
 0   SK_ID_PREV              1670214 non-null  int64
 1   SK_ID_CURR              1670214 non-null  int64
 2   NAME_CONTRACT_TYPE      1670214 non-null  object
 3   AMT_ANNUITY             1297979 non-null  float64
 4   AMT_APPLICATION         1670214 non-null  float64
 5   AMT_CREDIT              1670213 non-null  float64
 6   AMT_GOODS_PRICE         1284699 non-null  float64
 7   NAME_CASH_LOAN_PURPOSE  1670214 non-null  object
 8   NAME_CONTRACT_STATUS    1670214 non-null  object
 9   DAYS_DECISION           1670214 non-null  int64
 10  NAME_PAYMENT_TYPE       1670214 non-null  object
 11  CODE_REJECT_REASON      1670214 non-null  object
 12  NAME_CLIENT_TYPE        1670214 non-null  object
 13  NAME_GOODS_CATEGORY     1670214 non-null  object
 14  NAME_PORTFOLIO          1670214 non-null  object
 15  NAME_PRODUCT_TYPE       1670214 non-null  object
 16  CHANNEL_TYPE            1670214 non-null  object
 17  SELLERPLACE_AREA        1670214 non-null  int64
 18  NAME_SELLER_INDUSTRY    1670214 non-null  object
 19  CNT_PAYMENT             1297984 non-null  float64
 20  NAME_YIELD_GROUP        1670214 non-null  object
 21  PRODUCT_COMBINATION     1669868 non-null  object
dtypes: float64(5), int64(4), object(13)
memory usage: 280.3+ MB
```

Univariate Analysis – 'DAYS_DECISION_GRP'



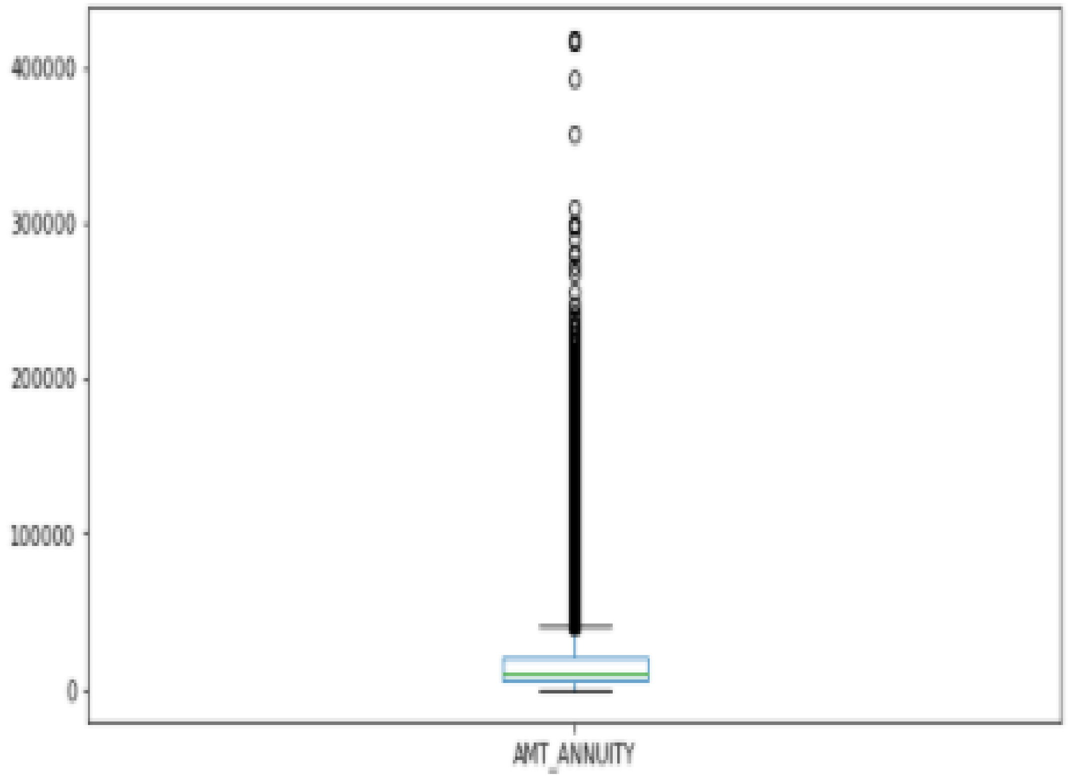
Inference:
38% of Loan Applicants applied for new loan within 400 days of previous loan

Null Value Imputation

AMT_GOODS_PRICE	23.08
CNT_PAYMENT	22.29
AMT_ANNUITY	22.29
PRODUCT_COMBINATION	0.02

Null Values has to be imputed for these columns.

Box Plot of 'AMT_ANNUITY' Column



We can see presence of huge outliers. Imputing with median and not mean, since mean is affected by outliers.

Imputation of Rest of the Columns

AMT_GOODS_PRICE → Replacing with mode since that is the most commonly occurring value

CNT_PAYMENT → For 'CNT_PAYMENT' null, 'NAME_CONTRACT_STATUS' is cancelled or refused or Unused. Thus, imputing with '0' to avoid skewness.

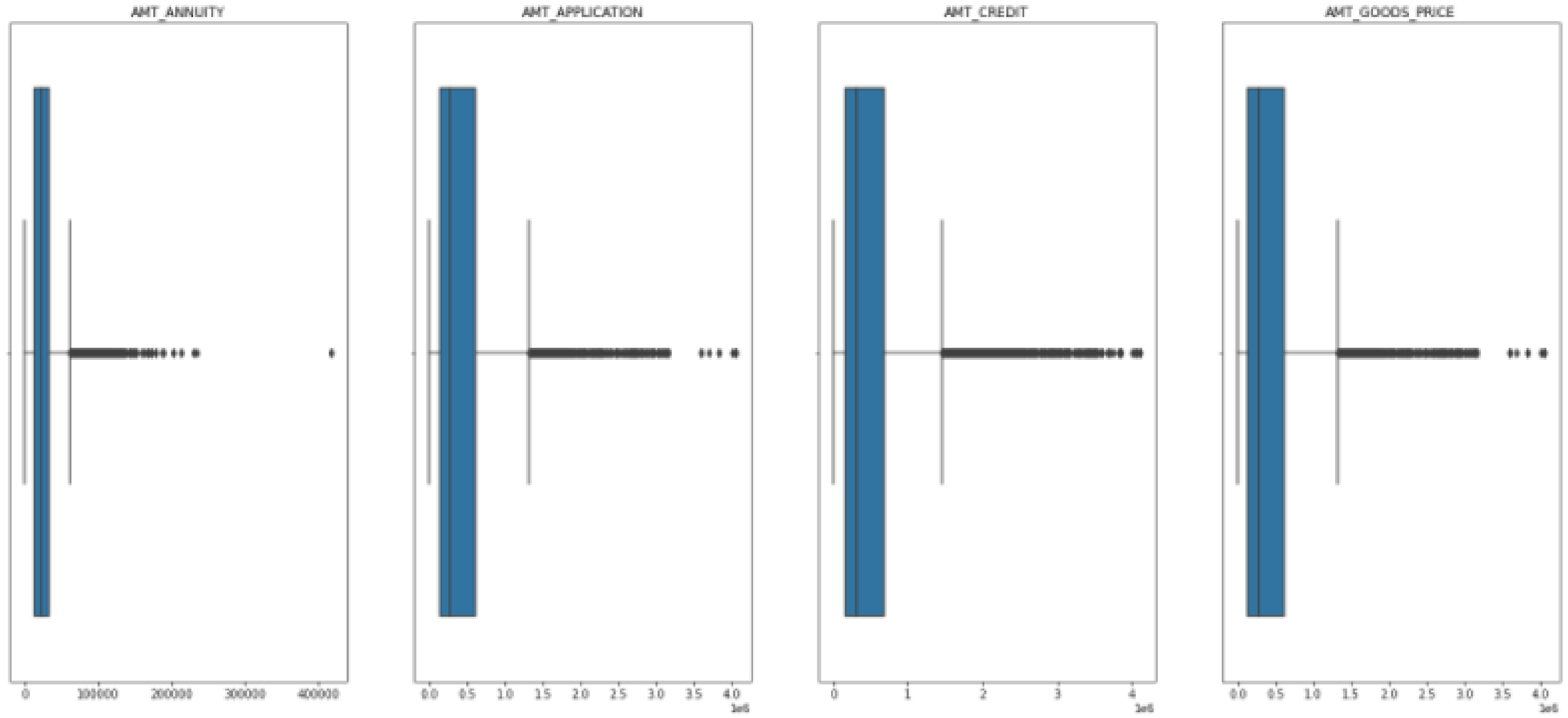
Removing ‘XNA’ and ‘XAP’ Records from ‘NAME_CASH_LOAN_PURPOSE’ Column

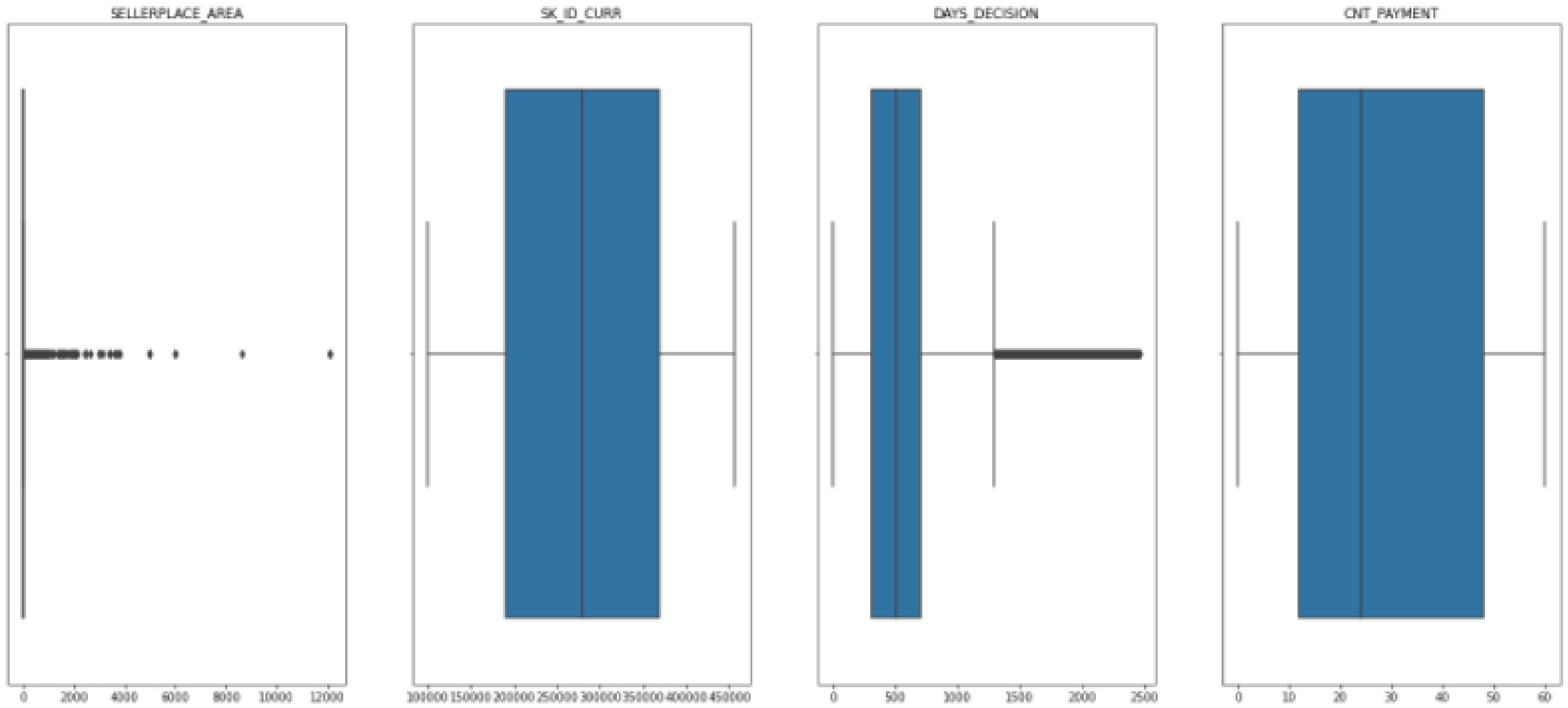
XAP	922661
XNA	677918

There were many garbage value which will impact our Analysis, hence we need to drop those.

Hence, post dropping the Dataframe dimensions became : (69635, 23)

Outlier Handling





Inference:
• CNT_PAYMENT has no outlier values.
• AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.
• DAYS_DECISION also has outliers.

Merged Data Analysis

Previous application and Current application data is than merged with 'SK_ID_CURR' to get desired Analysis. Dimension of merged dataframe : (59413, 74)

Data Structure of Merged Dataframe

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 59413 entries, 0 to 59412
Data columns (total 74 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   SK_ID_CURR                   59413 non-null  int64
 1   TARGET                       59413 non-null  int64
 2   NAME_CONTRACT_TYPE_x         59413 non-null  category
 3   CODE_GENDER                  59413 non-null  category
 4   FLAG_OWN_CAR                 59413 non-null  category
 5   FLAG_OWN_REALTY              59413 non-null  category
 6   CNT_CHILDREN                 59413 non-null  int64
 7   AMT_INCOME_TOTAL             59413 non-null  float64
 8   AMT_CREDIT_x                 59413 non-null  float64
 9   AMT_ANNUITY_x                59406 non-null  float64
 10  AMT_GOODS_PRICE_x            59413 non-null  float64
 11  NAME_TYPE_SUITE              59413 non-null  category
 12  NAME_INCOME_TYPE             59413 non-null  category
 13  NAME_EDUCATION_TYPE          59413 non-null  category
 14  NAME_FAMILY_STATUS           59413 non-null  category
 15  NAME_HOUSING_TYPE            59413 non-null  category
 16  REGION_POPULATION_RELATIVE   59413 non-null  float64
 17  DAYS_BIRTH                   59413 non-null  int64
 18  DAYS_EMPLOYED                59413 non-null  int64
 19  DAYS_REGISTRATION            59413 non-null  float64
 20  DAYS_ID_PUBLISH              59413 non-null  int64
 21  OCCUPATION_TYPE              59413 non-null  category
 22  CNT_FAM_MEMBERS              59413 non-null  float64
 23  REGION_RATING_CLIENT         59413 non-null  category
 24  REGION_RATING_CLIENT_W_CITY  59413 non-null  category
 25  WEEKDAY_APPR_PROCESS_START   59413 non-null  category
 26  HOUR_APPR_PROCESS_START      59413 non-null  int64
 27  REG_REGION_NOT_LIVE_REGION   59413 non-null  int64
 28  REG_REGION_NOT_WORK_REGION   59413 non-null  category
 29  LIVE_REGION_NOT_WORK_REGION  59413 non-null  category
 30  REG_CITY_NOT_LIVE_CITY       59413 non-null  category
 31  REG_CITY_NOT_WORK_CITY       59413 non-null  category
 32  LIVE_CITY_NOT_WORK_CITY      59413 non-null  category
 33  ORGANIZATION_TYPE            59413 non-null  category
 34  OBS_30_CNT_SOCIAL_CIRCLE     59413 non-null  float64
 35  DEF_30_CNT_SOCIAL_CIRCLE     59413 non-null  float64
 36  OBS_60_CNT_SOCIAL_CIRCLE     59413 non-null  float64
 37  DEF_60_CNT_SOCIAL_CIRCLE     59413 non-null  float64
```

```
 37  DEF_60_CNT_SOCIAL_CIRCLE     59413 non-null  float64
 38  DAYS_LAST_PHONE_CHANGE       59413 non-null  float64
 39  FLAG_DOCUMENT_3              59413 non-null  int64
 40  AMT_REQ_CREDIT_BUREAU_HOUR   59413 non-null  float64
 41  AMT_REQ_CREDIT_BUREAU_DAY    59413 non-null  float64
 42  AMT_REQ_CREDIT_BUREAU_WEEK   59413 non-null  float64
 43  AMT_REQ_CREDIT_BUREAU_MON    59413 non-null  float64
 44  AMT_REQ_CREDIT_BUREAU_QRT    59413 non-null  float64
 45  AMT_REQ_CREDIT_BUREAU_YEAR   59413 non-null  float64
 46  AMT_INCOME_RANGE             59380 non-null  category
 47  AMT_CREDIT_RANGE             59413 non-null  category
 48  AGE                          59413 non-null  int64
 49  AGE_GROUP                    59413 non-null  category
 50  YEARS_EMPLOYED               59413 non-null  int64
 51  EMPLOYMENT_YEAR              46747 non-null  category
 52  SK_ID_PREV                   59413 non-null  int64
 53  NAME_CONTRACT_TYPE_y         59413 non-null  category
 54  AMT_ANNUITY_y                59413 non-null  float64
 55  AMT_APPLICATION              59413 non-null  float64
 56  AMT_CREDIT_y                 59413 non-null  float64
 57  AMT_GOODS_PRICE_y            59413 non-null  float64
 58  NAME_CASH_LOAN_PURPOSE       59413 non-null  category
 59  NAME_CONTRACT_STATUS         59413 non-null  category
 60  DAYS_DECISION                59413 non-null  int64
 61  NAME_PAYMENT_TYPE            59413 non-null  category
 62  CODE_REJECT_REASON           59413 non-null  category
 63  NAME_CLIENT_TYPE             59413 non-null  category
 64  NAME_GOODS_CATEGORY          59413 non-null  category
 65  NAME_PORTFOLIO               59413 non-null  category
 66  NAME_PRODUCT_TYPE            59413 non-null  category
 67  CHANNEL_TYPE                 59413 non-null  category
 68  SELLERPLACE_AREA             59413 non-null  int64
 69  NAME_SELLER_INDUSTRY         59413 non-null  category
 70  CNT_PAYMENT                  59413 non-null  float64
 71  NAME_YIELD_GROUP             59413 non-null  category
 72  PRODUCT_COMBINATION          59413 non-null  category
 73  DAYS_DECISION_GRP            59413 non-null  category
dtypes: category(37), float64(23), int64(14)
memory usage: 19.3 MB
```

Data division based on Target Column

In [129]:
```
# Splitting 'df_merged' dataframe into two dataframes based on 'TARGET' values

df_merged_repayers = df_merged[df_merged['TARGET']==0]
df_merged_defaulters = df_merged[df_merged['TARGET']==1]
```

In [130]:
```
# Reading the first 3 lines from the dataframe 'df_merged_repayers'

df_merged_repayers.head(3)
```

Out[130]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE_x	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_C
0	100034	0	Revolving loans	M	N	Y	0	0.900	
1	100035	0	Cash loans	F	N	Y	0	2.925	
2	100039	0	Cash loans	M	Y	N	1	3.600	

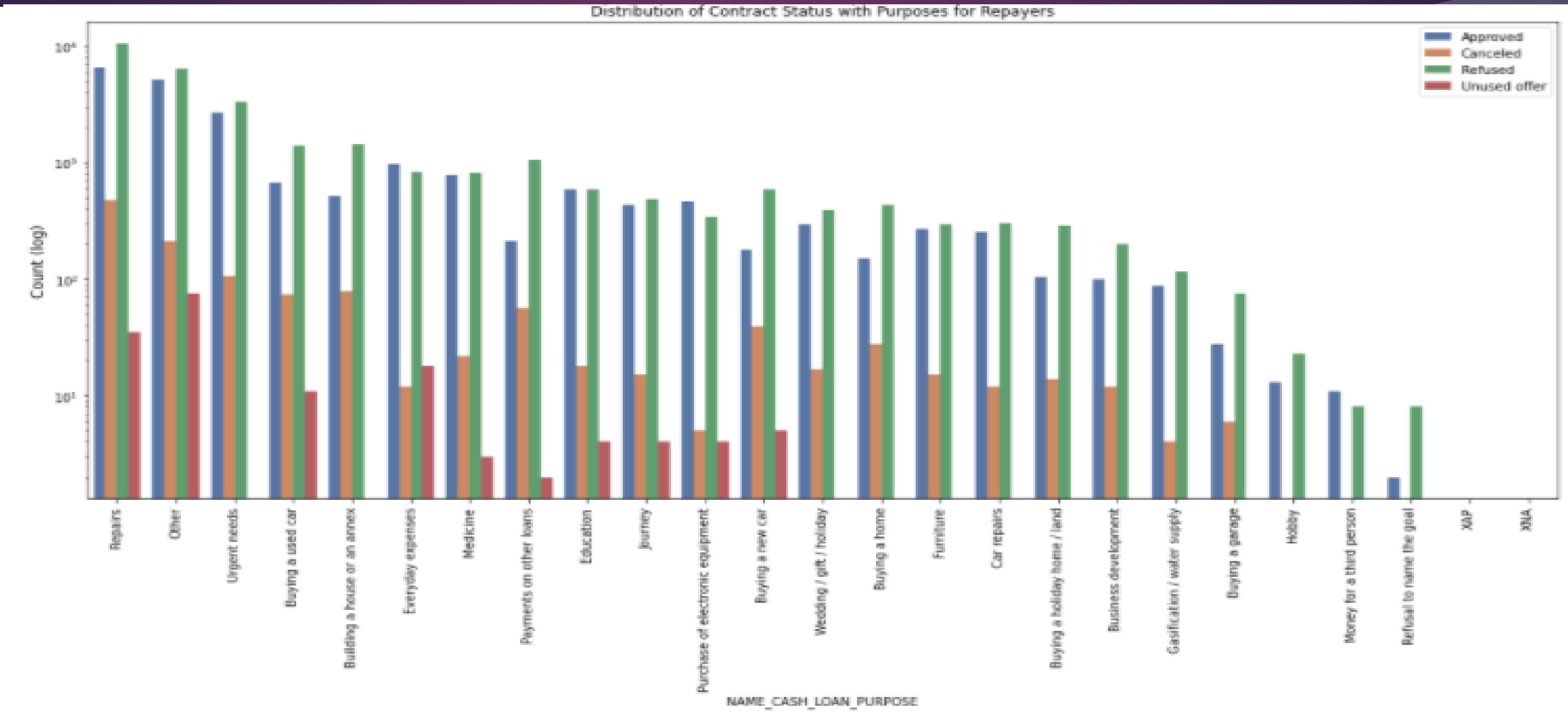
In [131]:
```
# Reading the first 3 lines from the dataframe 'df_merged_defaulters'

df_merged_defaulters.head(3)
```

Out[131]:

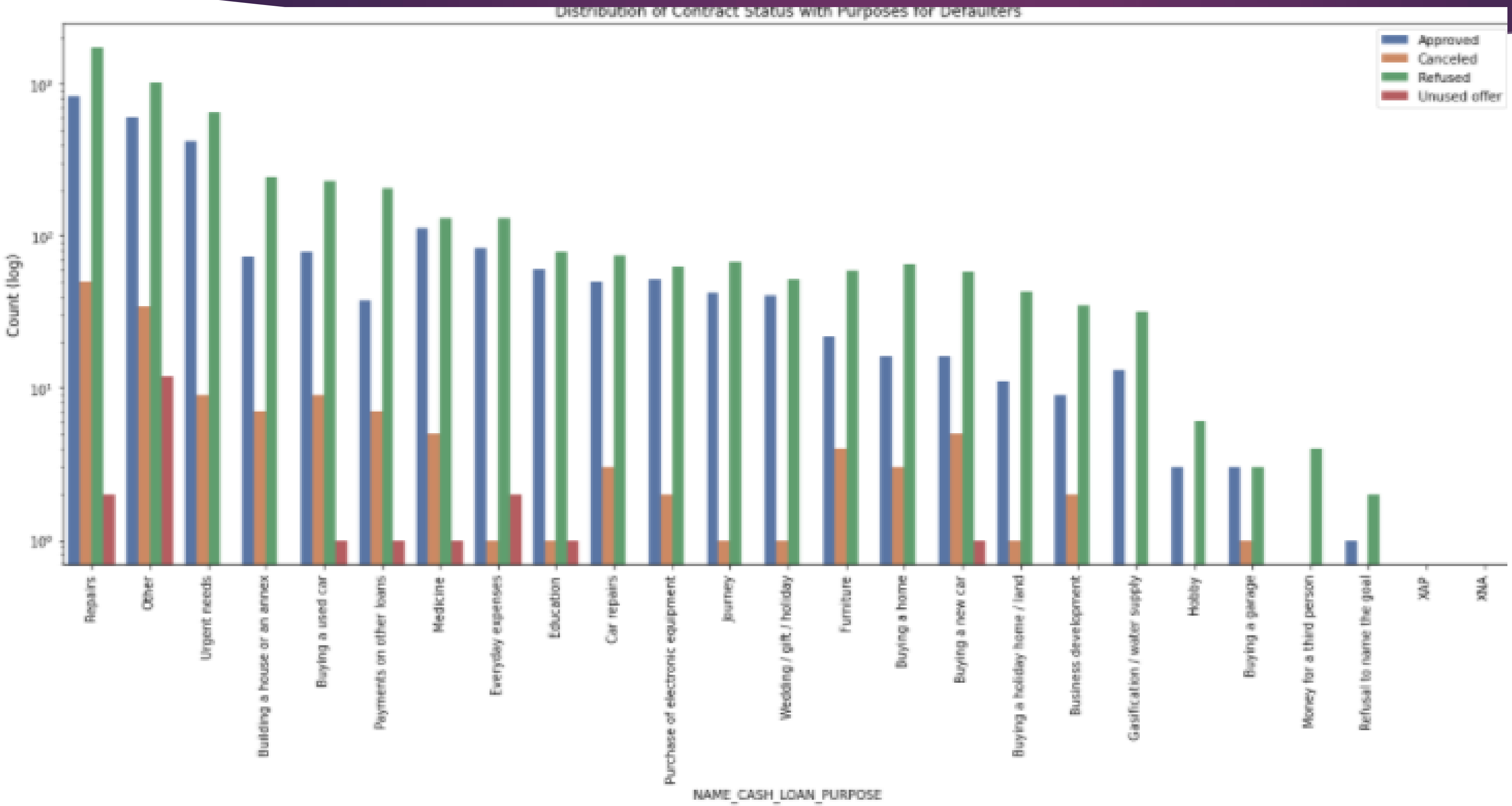
	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE_x	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_
36	100301	1	Cash loans	M	N	Y	1	1.125	
86	100547	1	Cash loans	M	Y	N	0	2.115	
87	100547	1	Cash loans	M	Y	N	0	2.115	

Univariate Analysis – Merged Dataframe

Analysis of ‘NAME_CASH_LOAN_PURPOSE’ Column for Non-Defaulters



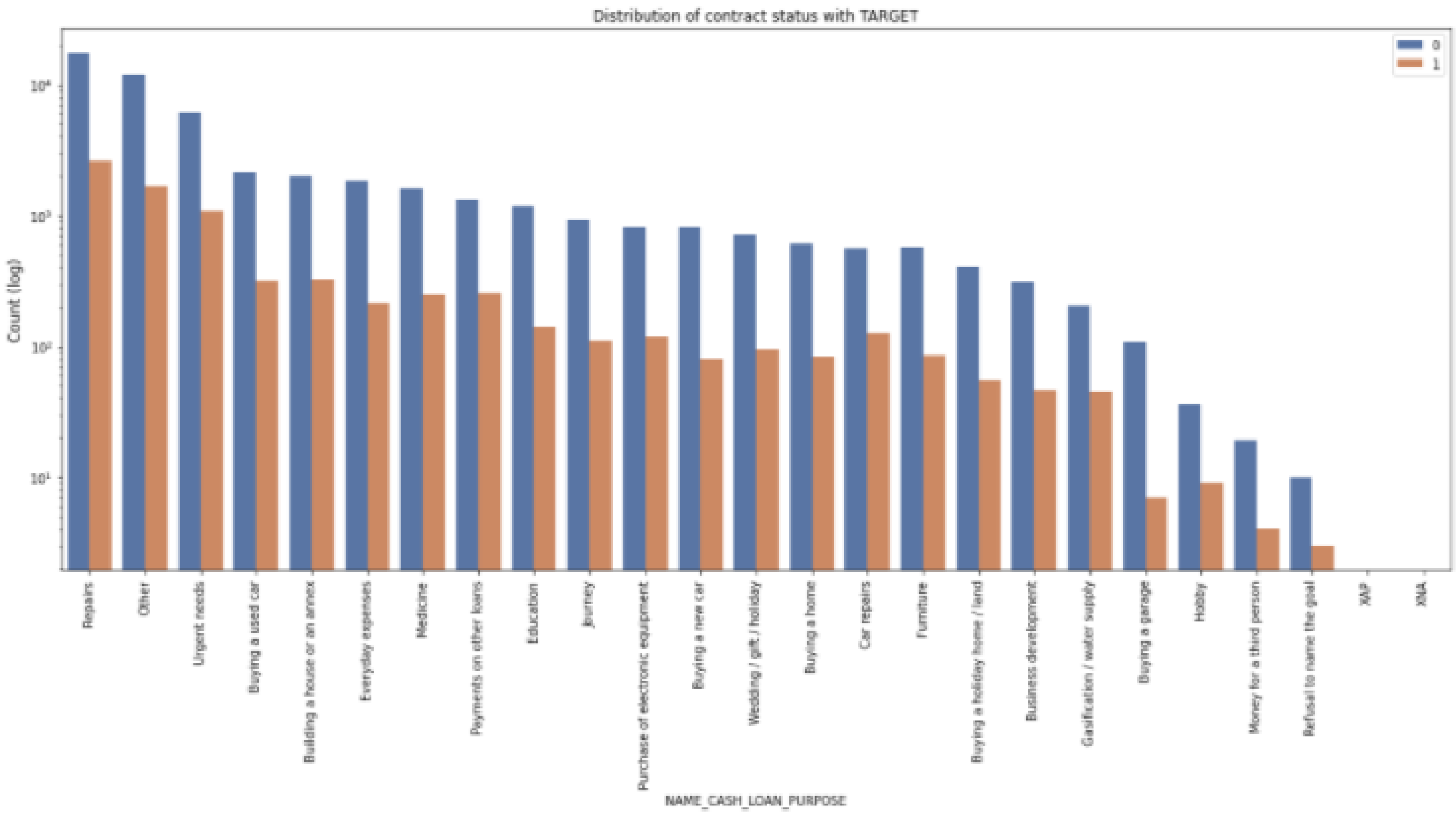
Analysis of ‘NAME_CASH_LOAN_PURPOSE’ Column for Defaulters



Inference:

- Most rejection of loans came for purpose 'Repair', 'Other' and 'Urgent needs'.
- For 'Education' purpose we have equal number of 'Approval' and 'Rejection'.
- Buying a new car is having significant higher rejection rate than approval rate.

Analysis of ‘NAME_CASH_LOAN_PURPOSE’ vs. ‘TARGET’

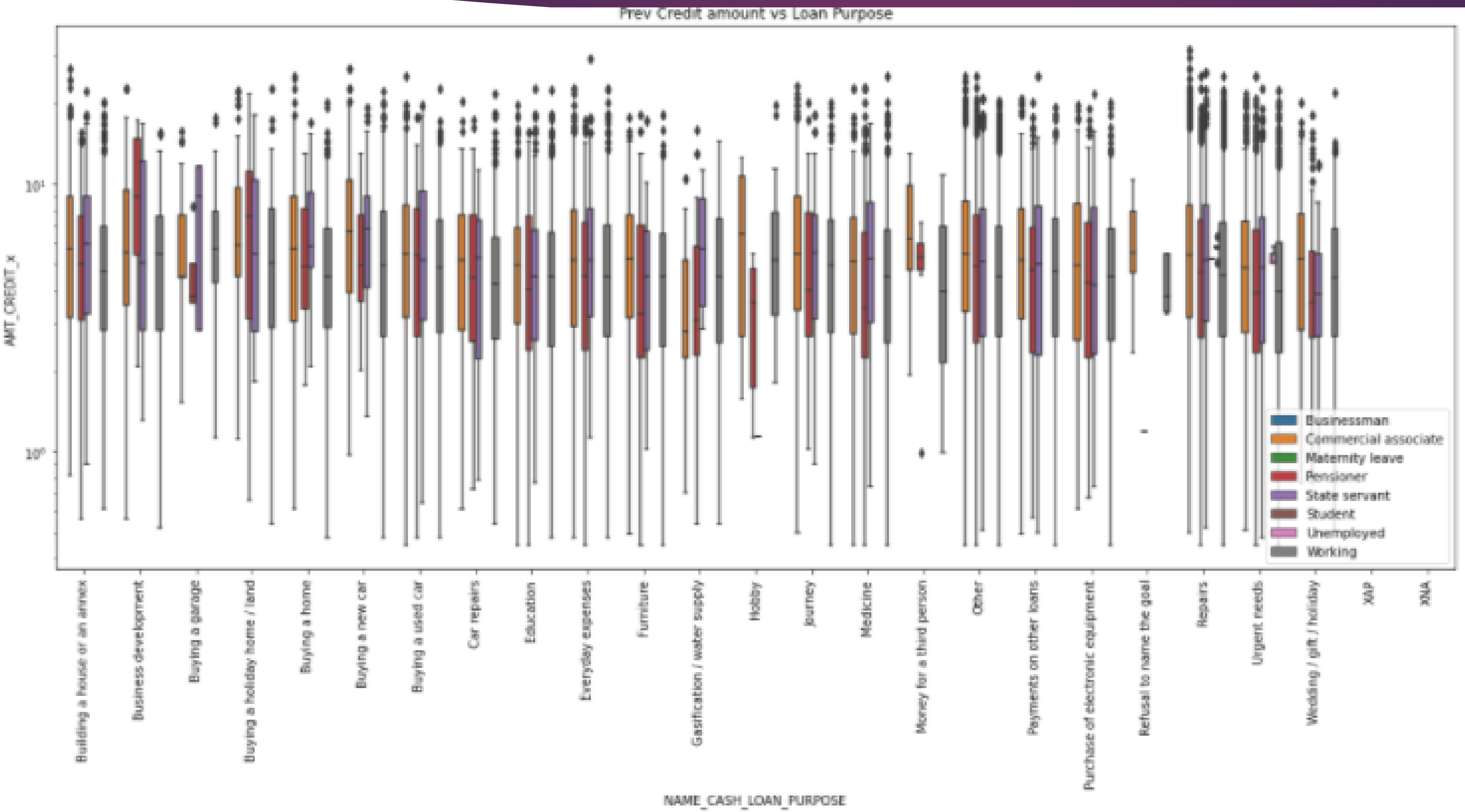


Inference:

• Loan purpose with 'Repairs' has a high default rate.
Loan purpose with 'Buying a garage' has lower default rate.
• Loan purpose with 'Business development' has lower default rate.
• Loan purpose with 'Buying a new car' has lower default rate.
Loan purpose with 'Education' has lower default rate.

Bivariate Analysis – Merged Dataframe

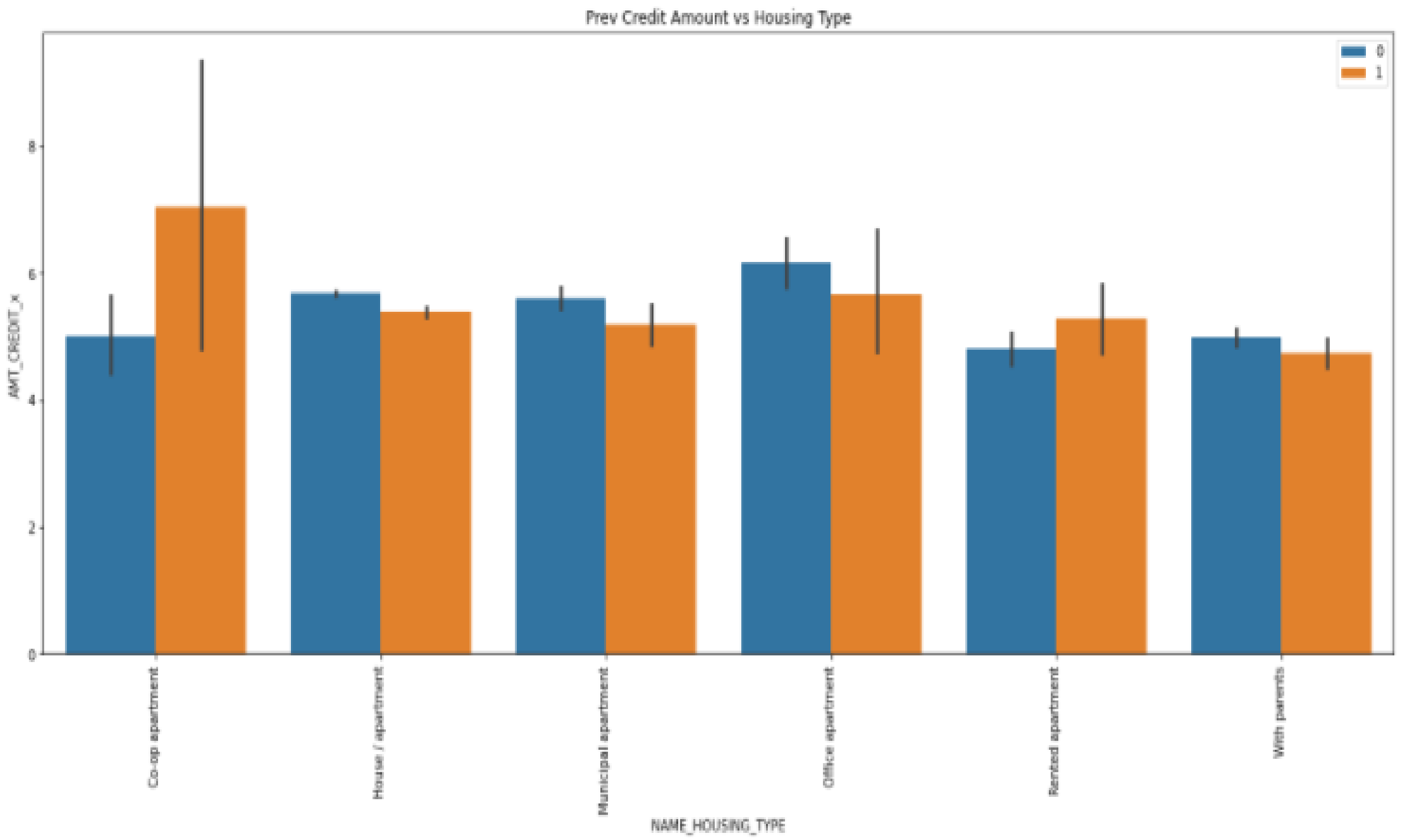
‘NAME_CASH_LOAN_PURPOSE’ vs. ‘AMT_CREDIT_x’ Columns



Inference:

- Money for third person or a Hobby is having less credits.
- Income type 'State servants' have a significant amount of credit.
- Credit amtt of 'Buying a holiday home/land', 'Buying a new car', 'Building a house' is higher.

‘NAME_HOUSING_TYPE’ vs. ‘AMT_CREDIT_x’ Columns



Inference:

- Co-op apartment has higher default rate.
- Rented apartment also has slightly higher default rate.
- Office apartment has higher repay rate.

Conclusions

Decisive Factors Whether Applicants will be Repayers

• Students and Businessmen have no defaults.

•Applicants having Education Type as 'Academic Degree' has less default rate compared to others.

•Applicants with 'Trade Type 4' Organization Type has less default rate compared to others.

•Applicants above age of 50 have a low probability of defaulting.

•Applicants whose Income is between 700k-800k are less likely to default.

•Applicants with more than 40+ years of experience are less likely to default.

•Applicants having 0-2 children have low probability of defaulting.

•Applicants from Housing Type 'With Parents' have a low probability of defaulting.

Decisive Factors Whether Applicants will be Defaulters

- Men have relatively higher default rate.
- Applicants whose Education Type is 'Lower Secondary' or 'Secondary' defaults a lot.
- Applicants who are either 'Unemployed' or on 'Maternity Leave' have a high default rate.
- Applicants living in 'Rating 3' have high default rate.
- Applicants who are 'Single' or had 'Civil Marriage' defaults a lot.
- Applicants whose Occupation types are 'Low-Skill Labourers','Drivers','Waiters','Security Staffs' have a high default rate.
- Applicants in Age Group 20-40 have a high default rate.
- Applicants having less than 5 years experience have a high default rate.
- Applicants having more than 9 childrens have a high default rate.
- Applicants with Loan Purpose 'Repair' have high default rate.

Thank you