Project 6 : By Ramana Bansal

Bank Loan Case Study

Colab notebook Link:
https://colab.research.google.com/drive/1gdARCHcwWReZ1gnJfT9DFgsUbz-9wL8f?usp=sharing

Video Link:
https://drive.google.com/file/d/1F2NBCVzVq82tYzOXjDrk0B7p-JP-zbtQ/view?usp=sharing

Project Description: The project deals with risk analytics related to loan applications in a bank. The aim of the project is to use EDA to identify the factors and patterns which may indicate that an applicant might have difficulty in loan payment and use this to identify the applications that should be approved or not, in order to reduce loan defaults.

Problem Statement: Analyzing provided data to predict whether an applicant might default in loan payment or not.

Data Sets:

- application_data : Data regarding the current applications and applicants' details.
- previous_application_data : Data regarding the previous applications of applicants.

Analysis Approach: The two datasets were initially processed and analyzed separately, and then the data was merged. The following steps were taken:

1. Importing required libraries : numpy, pandas, matplotlib and seaborn.
2. Mounting Google Drive : Since Google Colab allocates fresh RAM for every session, files need to be uploaded for every session. With heavier files, it's better to simply connect Colab to Drive for easier access to files.
3. Working with application_data file and previous_application_data:
    - Understanding data
    - Removing columns with high null data
    - Removing duplicates
    - Checking data imbalance, before and after merging

- Dividing data into Categorical, Discrete and Numerical columns and working on them separately for
    a. Working on missing or unknown values
    b. Changing datatypes
    c. Treating Outliers
    d. Univariate Analysis
    e. Bivariate Analysis
    f. Finding Correlation
    g. Visualization

Tech-Stack Used: The data was processed and analyzed using Google Colab.

Learning Insights: The analysis highlighted various features which might aid in predicting whether an applicant might default or not. It also helped in understanding the type of loan applications and the type of loan applicants a bank gets.

The project gave me an opportunity to revisit Python as well as learn some of its required libraries. It also helped me to understand the work and approach required to work with large amount of data. However, it was a little difficult for me to draw insights from data. The project highlighted the need to work on the same.

Missing Data

A. Identify the missing data and use appropriate method to deal with it.

First, we checked the percentage of null values for each column. For application_data, the columns with null percentage > 40% were dropped. For previous_application_data, the columns with null percentage > 50% were dropped.

Then, we checked the columns for XNA values in Categorical columns. If less in number, these were replaced by mode, else, they were replaced by NAN. The percentage of null values was then rechecked. The columns with XNA>50% were also removed.

For numerical columns with few missing values, the outliers were checked. In case of presence of outliers, the null values were imputed with median. If there were no outliers, the null values were replaced by mean. If the number of missing values was high, no imputation was made.

Outliers

B. Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier.

The outliers in numerical columns were checked using Box-plot. The values falling above or below the IQR values were considered outliers. There were no outliers below the lower bound. The outliers lying above IQR were capped with 99 percentile values instead of being removed.

For some of DAYS columns, an error value of 365243 (~100 years) was observed. This value was NaN. The DAYS columns were then converted into Years and stored in dataframes.

Data Imbalance

C. Identify if there is data imbalance in the data. Find the ratio of data imbalance.

Since the major aim of study was to differentiate between people with payment difficulties (defaulters) and non-defaulters, the TARGET column was used to check data imbalance. The column had following two values:
0: Applicants with no payment difficulties (Non-Defaulters)
1: Applicants with payment difficulties (Defaulters)

The ratio between the above two values was found to check Data Imbalance.

Working on application_data

Description

The dataframe app_data has 122 columns and 307511 rows. There are 65 columns with float datatype, 41 with int and 16 with object datatype.

```
app_data.info()
```

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

COLUMNS:

```
['SK_ID_CURR',
 'TARGET',
 'NAME_CONTRACT_TYPE',
 'CODE_GENDER',
 'FLAG_OWN_CAR',
 'FLAG_OWN_REALTY',
 'CNT_CHILDREN',
 'AMT_INCOME_TOTAL',
 'AMT_CREDIT',
 'AMT_ANNUITY',
 'AMT_GOODS_PRICE',
 'NAME_TYPE_SUITE',
 'NAME_INCOME_TYPE',
 'NAME_EDUCATION_TYPE',
 'NAME_FAMILY_STATUS',
 'NAME_HOUSING_TYPE',
 'REGION_POPULATION_RELATIVE',
 'DAYS_BIRTH',
 'DAYS_EMPLOYED',
 'DAYS_REGISTRATION',
 'DAYS_ID_PUBLISH',
 'OWN_CAR_AGE',
 'FLAG_MOBIL',
 'FLAG_EMP_PHONE',
 'FLAG_WORK_PHONE',
 'FLAG_CONT_MOBILE',
 'FLAG_PHONE',
 'FLAG_EMAIL',
 'OCCUPATION_TYPE',
 'CNT_FAM_MEMBERS',
 'REGION_RATING_CLIENT',
 'REGION_RATING_CLIENT_W_CITY',
 'WEEKDAY_APPR_PROCESS_START',
 'HOUR_APPR_PROCESS_START',
 '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',
 'ORGANIZATION_TYPE',
 'EXT_SOURCE_1',
 'EXT_SOURCE_2',
 'EXT_SOURCE_3',
```

```
 'EXT_SOURCE_3',
 'APARTMENTS_AVG',
 'BASEMENTAREA_AVG',
 'YEARS_BEGINEXPLUATATION_AVG',
 'YEARS_BUILD_AVG',
 'COMMONAREA_AVG',
 'ELEVATORS_AVG',
 'ENTRANCES_AVG',
 'FLOORSMAX_AVG',
 'FLOORSMIN_AVG',
 'LANDAREA_AVG',
 'LIVINGAPARTMENTS_AVG',
 'LIVINGAREA_AVG',
 'NONLIVINGAPARTMENTS_AVG',
 'NONLIVINGAREA_AVG',
 'APARTMENTS_MODE',
 'BASEMENTAREA_MODE',
 'YEARS_BEGINEXPLUATATION_MODE',
 'YEARS_BUILD_MODE',
 'COMMONAREA_MODE',
 'ELEVATORS_MODE',
 'ENTRANCES_MODE',
 'FLOORSMAX_MODE',
 'FLOORSMIN_MODE',
 'LANDAREA_MODE',
 'LIVINGAPARTMENTS_MODE',
 'LIVINGAREA_MODE',
 'NONLIVINGAPARTMENTS_MODE',
 'NONLIVINGAREA_MODE',
 'APARTMENTS_MEDI',
 'BASEMENTAREA_MEDI',
 'YEARS_BEGINEXPLUATATION_MEDI',
 'YEARS_BUILD_MEDI',
 'COMMONAREA_MEDI',
 'ELEVATORS_MEDI',
 'ENTRANCES_MEDI',
 'FLOORSMAX_MEDI',
 'FLOORSMIN_MEDI',
 'LANDAREA_MEDI',
 'LIVINGAPARTMENTS_MEDI',
 'LIVINGAREA_MEDI',
 'NONLIVINGAPARTMENTS_MEDI',
 'NONLIVINGAREA_MEDI',
 'FONDKAPREMONT_MODE',
```

```
 'FONDKAPREMONT_MODE',
 'HOUSETYPE_MODE',
 'TOTALAREA_MODE',
 'WALLSMATERIAL_MODE',
 'EMERGENCYSTATE_MODE',
 'OBS_30_CNT_SOCIAL_CIRCLE',
 'DEF_30_CNT_SOCIAL_CIRCLE',
 'OBS_60_CNT_SOCIAL_CIRCLE',
 'DEF_60_CNT_SOCIAL_CIRCLE',
 'DAYS_LAST_PHONE_CHANGE',
 'FLAG_DOCUMENT_2',
 'FLAG_DOCUMENT_3',
 'FLAG_DOCUMENT_4',
 'FLAG_DOCUMENT_5',
 'FLAG_DOCUMENT_6',
 'FLAG_DOCUMENT_7',
 'FLAG_DOCUMENT_8',
 'FLAG_DOCUMENT_9',
 'FLAG_DOCUMENT_10',
 'FLAG_DOCUMENT_11',
 'FLAG_DOCUMENT_12',
 'FLAG_DOCUMENT_13',
 'FLAG_DOCUMENT_14',
 'FLAG_DOCUMENT_15',
 'FLAG_DOCUMENT_16',
 'FLAG_DOCUMENT_17',
 'FLAG_DOCUMENT_18',
 'FLAG_DOCUMENT_19',
 'FLAG_DOCUMENT_20',
 'FLAG_DOCUMENT_21',
 'AMT_REQ_CREDIT_BUREAU_HOUR',
 'AMT_REQ_CREDIT_BUREAU_DAY',
 'AMT_REQ_CREDIT_BUREAU_WEEK',
 'AMT_REQ_CREDIT_BUREAU_MON',
 'AMT_REQ_CREDIT_BUREAU_QRT',
 'AMT_REQ_CREDIT_BUREAU_YEAR']
```

Irrelevant Columns

The following **columns with null values > 40%** were removed.

```
OWN_CAR_AGE                         65.990810
EXT_SOURCE_1                        56.381073
APARTMENTS_AVG                      50.749729
BASEMENTAREA_AVG                    58.515956
YEARS_BEGINEXPLUATATION_AVG         48.781019
YEARS_BUILD_AVG                     66.497784
COMMONAREA_AVG                      69.872297
ELEVATORS_AVG                       53.295980
ENTRANCES_AVG                       50.348768
FLOORSMAX_AVG                       49.760822
FLOORSMIN_AVG                       67.848630
LANDAREA_AVG                        59.376738
LIVINGAPARTMENTS_AVG                68.354953
LIVINGAREA_AVG                      50.193326
NONLIVINGAPARTMENTS_AVG             69.432963
NONLIVINGAREA_AVG                   55.179164
APARTMENTS_MODE                     50.749729
BASEMENTAREA_MODE                   58.515956
YEARS_BEGINEXPLUATATION_MODE        48.781019
YEARS_BUILD_MODE                    66.497784
COMMONAREA_MODE                     69.872297
ELEVATORS_MODE                      53.295980
ENTRANCES_MODE                      50.348768
FLOORSMAX_MODE                      49.760822
FLOORSMIN_MODE                      67.848630
LANDAREA_MODE                       59.376738
LIVINGAPARTMENTS_MODE               68.354953
LIVINGAREA_MODE                     50.193326
NONLIVINGAPARTMENTS_MODE            69.432963
NONLIVINGAREA_MODE                  55.179164
APARTMENTS_MEDI                     50.749729
BASEMENTAREA_MEDI                   58.515956
YEARS_BEGINEXPLUATATION_MEDI        48.781019
YEARS_BUILD_MEDI                    66.497784
COMMONAREA_MEDI                     69.872297
ELEVATORS_MEDI                      53.295980
ENTRANCES_MEDI                      50.348768
FLOORSMAX_MEDI                      49.760822
FLOORSMIN_MEDI                      67.848630
LANDAREA_MEDI                       59.376738
LIVINGAPARTMENTS_MEDI               68.354953
LIVINGAREA_MEDI                     50.193326
NONLIVINGAPARTMENTS_MEDI            69.432963
NONLIVINGAREA_MEDI                  55.179164
```

```
NONLIVINGAPARTMENTS_MEDI            69.432963
NONLIVINGAREA_MEDI                  55.179164
FONDKAPREMONT_MODE                  68.386172
HOUSETYPE_MODE                      50.176091
TOTALAREA_MODE                      48.268517
WALLSMATERIAL_MODE                  50.840783
EMERGENCYSTATE_MODE                 47.398304
dtype: float64
```

43 columns were removed from app_data and resulting data was stored in df1. Df1 has 73 columns.

Duplicates

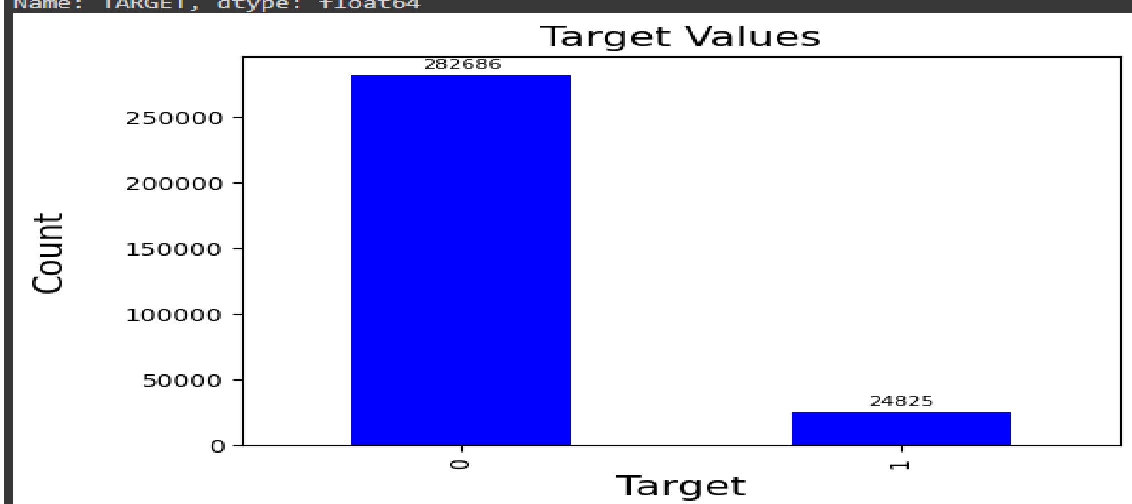
No duplicates were found in df1.

```
[15] duplicate1 = df1[df1.duplicated()]

     print("Duplicate Rows :")
     duplicate1

Duplicate Rows :
   SK_ID_CURR  TARGET  NAME_CONTRACT_TYPE  CODE_GENDER  FLAG_OWN_CAR  FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL
```

Data Imbalance

```
Since the major aim of study is to look into applicants with paying
difficulties, the target column will be used to check for data imbalance.
# Value = 0 indicates No Payment Difficulties (Non-Defaulters).
# Value = 1 indicates Payment Difficulties (Defaulters).
```

```
0 : Non-Defaulters
1 : Defaulters

0    91.927118
1     8.072882
Name: TARGET, dtype: float64
```

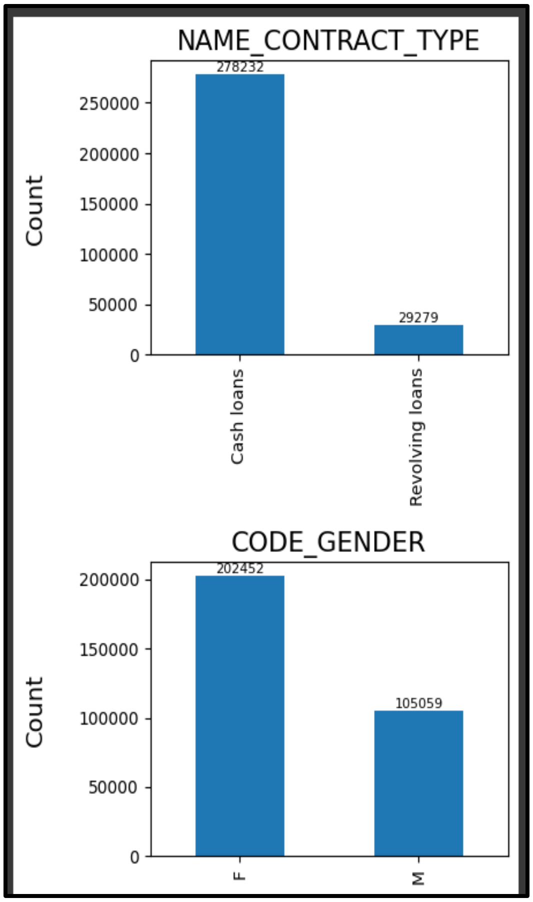


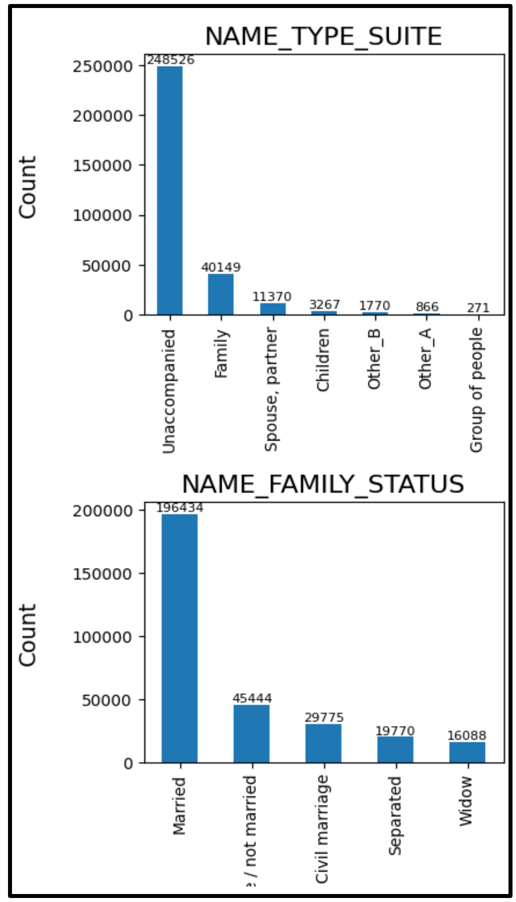
Data Imbalance ratio of 23:2 indicates the number/data of non-defaulters is much higher than that of defaulters.

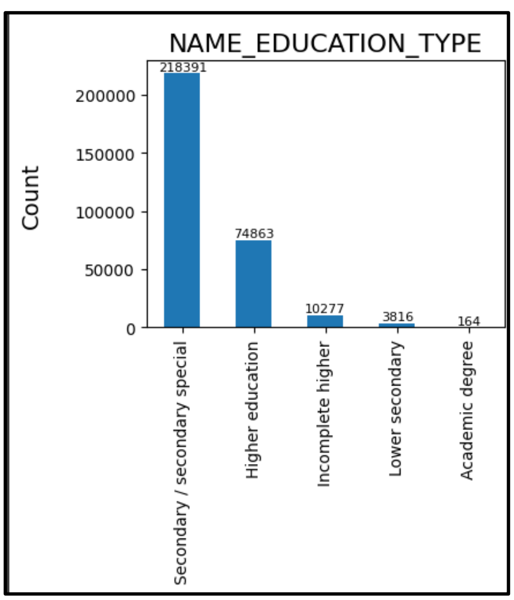
Univariate Analysis

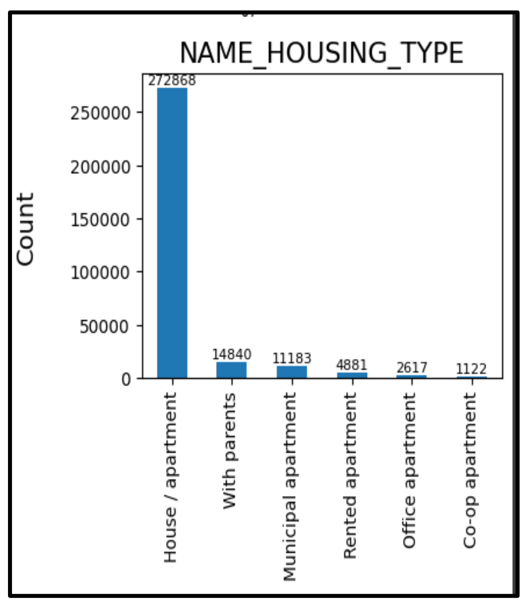
Categorical Columns

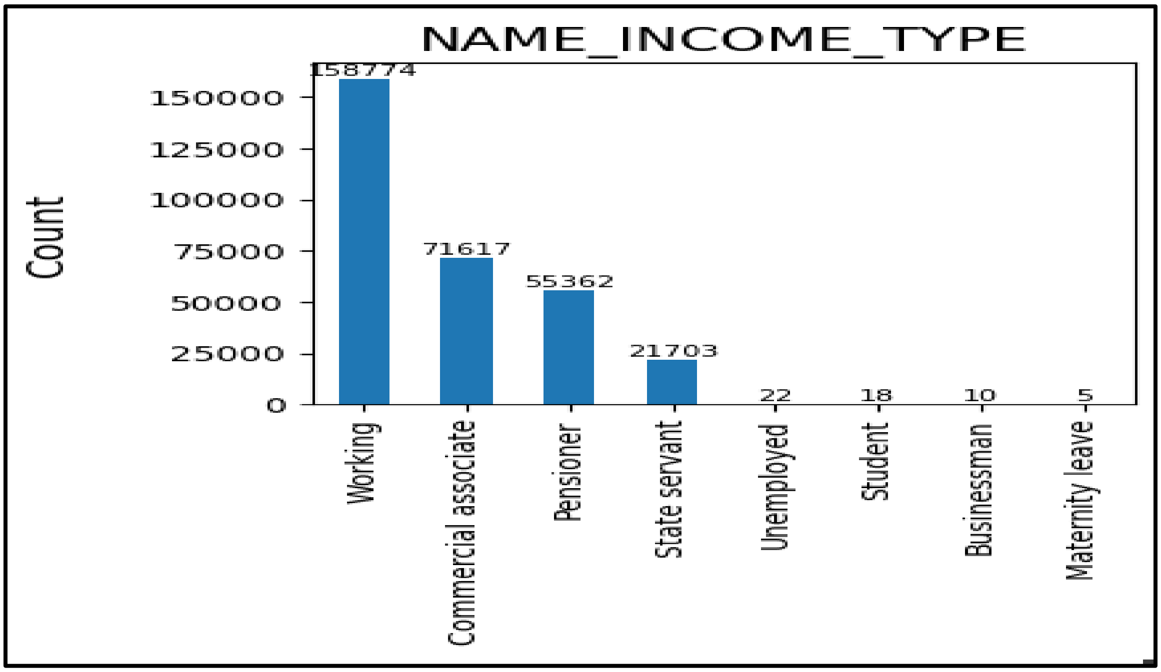
1. About 90.5% of loans were Cash loans while only 9.5% were Revolving loans.
2. The number of female applicants (65%) was almost double of male applicants (35%).
3. Most of the applicants had Secondary or Secondary Special education (71%), followed by Higher education (24%). The least number of applicants were from people with an academic degree.
4. Most of the applicants lived unaccompanied (80%). About 13% lived with their families.
5. 63% of applicants were married, 14% were single, 9.6% had civil marriage, 6.4% were separated and about 5% were widows.
6. 88% of the applicants lived in a house or apartment.
7. About 70% of applicants owned realty while 30% didn't.
8. About 34% of applicants owned cars while 64% didn't.
9. Most of the applicants were Working or Commercial associates. Businessmen and people on maternity leave had the least number of applications.
10. The maximum number of applicants was of laborers (17%), followed by sales Staff (10%).
11. People from Business Entity type 3 (22%) applied the most for loan, followed by self-employed people (12%).
12. For most of applicants, registration region was neither work nor live region.
13. Most of the applicants had provided their mobile phone numbers, work phone numbers and email-ids. Moreover, for most of the applicants the number was found to be reachable.
14. Among required documents, only Document 3 was provided by 70% of the applicants, while other documents were not provided by most.


NAME_CONTRACT_TYPE
Count
278232
29279
Cash loans
Revolving loans
CODE_GENDER
Count
202452
105059
F
M

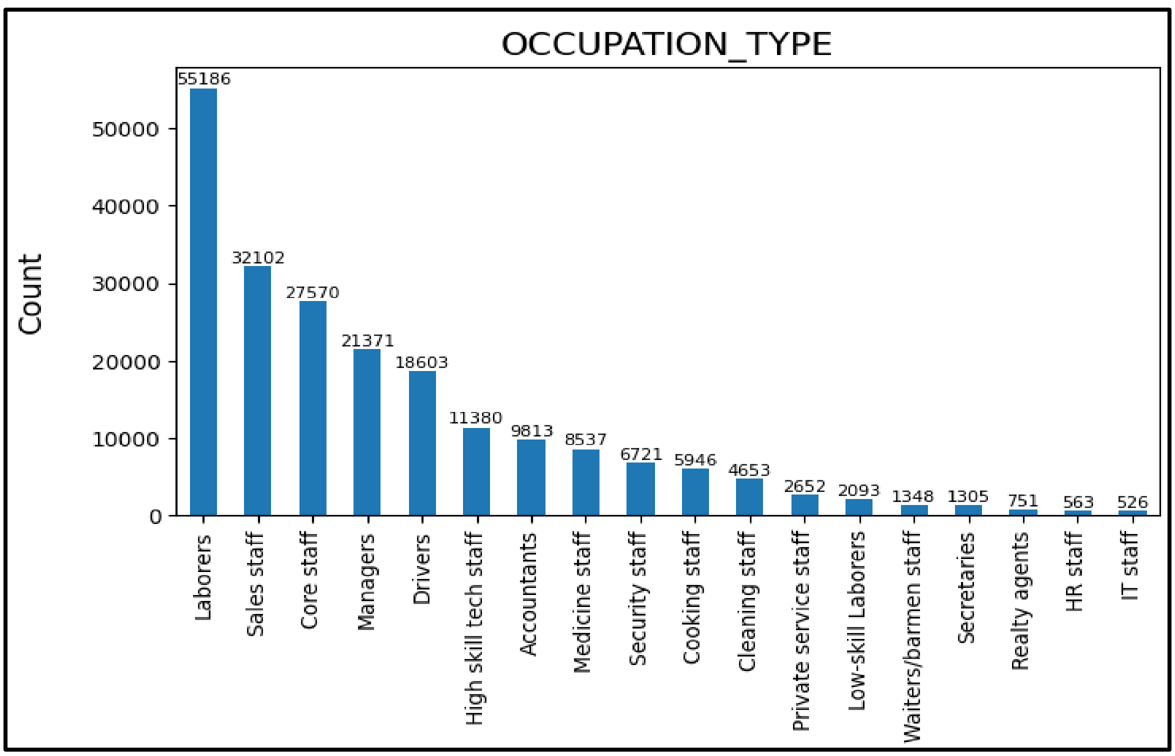

NAME_TYPE_SUITE
Count
248526
40149
11370
3267
1770
866
271
Unaccompanied
Family
Spouse, partner
Children
Other_B
Other_A
Group of people
NAME_FAMILY_STATUS
Count
196434
45444
29775
19770
16088
Married
Civil marriage
Separated
Widow


NAME_EDUCATION_TYPE
Count
218391
74863
10277
3816
164
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree


NAME_HOUSING_TYPE
Count
272868
14840
11183
4881
2617
1122
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment
Co-op apartment



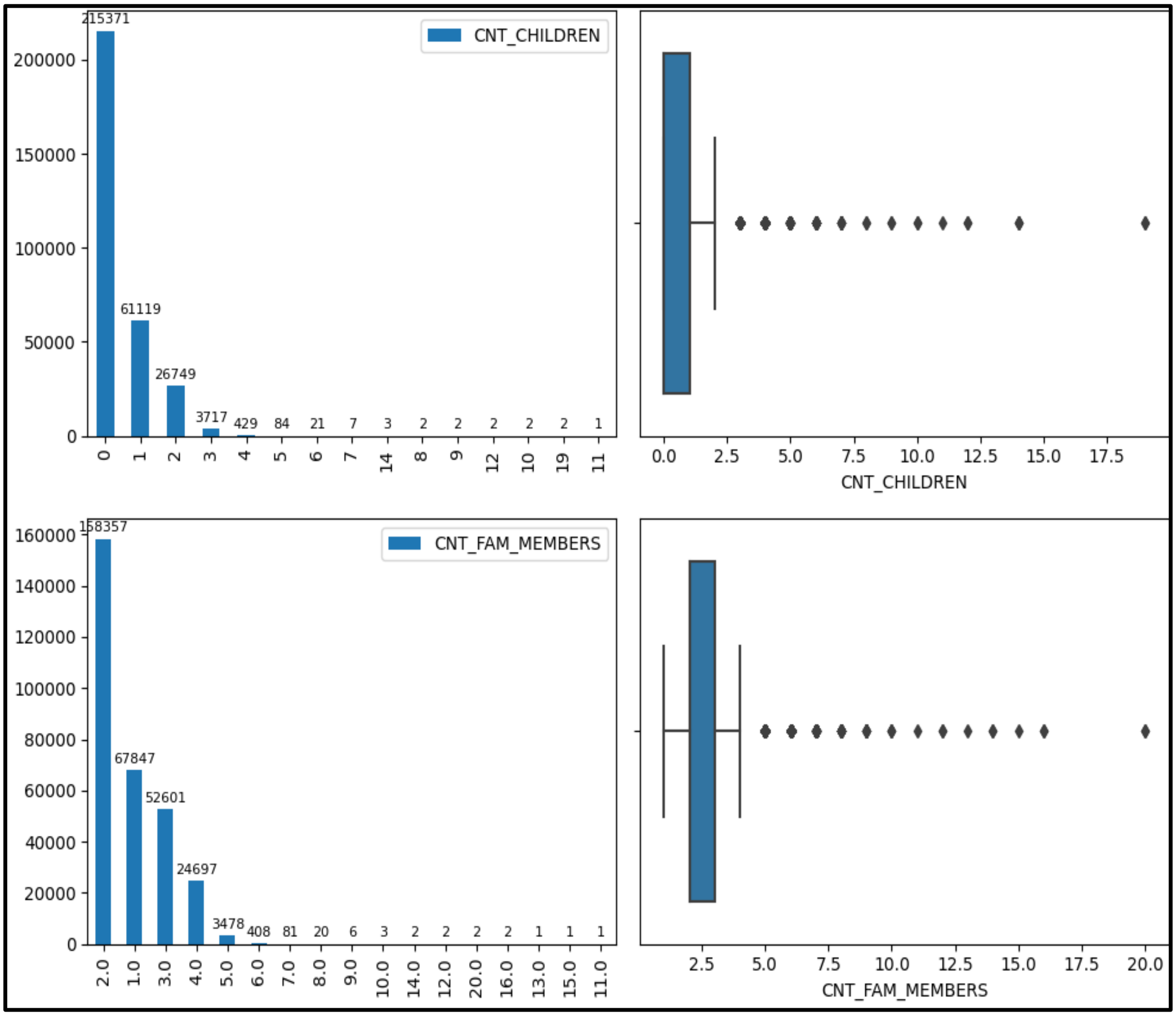
NAME_INCOME_TYPE
Count
150000
125000
100000
75000
50000
25000
0
158774
71617
55362
21703
22
18
10
5
Working
Commercial associate
Pensioner
State servant
Unemployed
Student
Businessman
Maternity leave

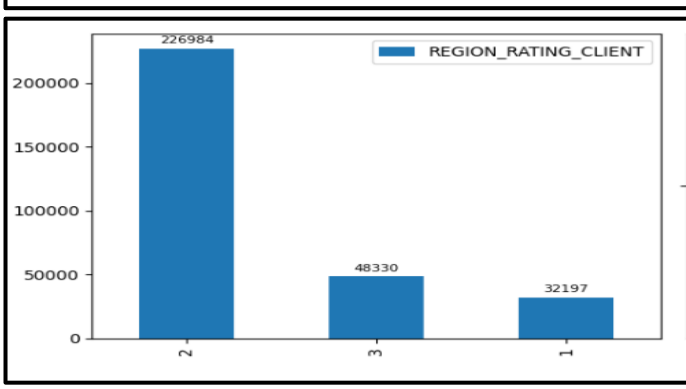


OCCUPATION_TYPE
Count
50000
40000
30000
20000
10000
0
55186
32102
27570
21371
18603
11380
9813
8537
6721
5946
4653
2652
2093
1348
1305
751
563
526
Laborers
Sales staff
Core staff
Managers
Drivers
High skill tech staff
Accountants
Medicine staff
Security staff
Cooking staff
Cleaning staff
Private service staff
Low-skill Laborers
Waiters/barmen staff
Secretaries
Realty agents
HR staff
IT staff

Discrete Columns

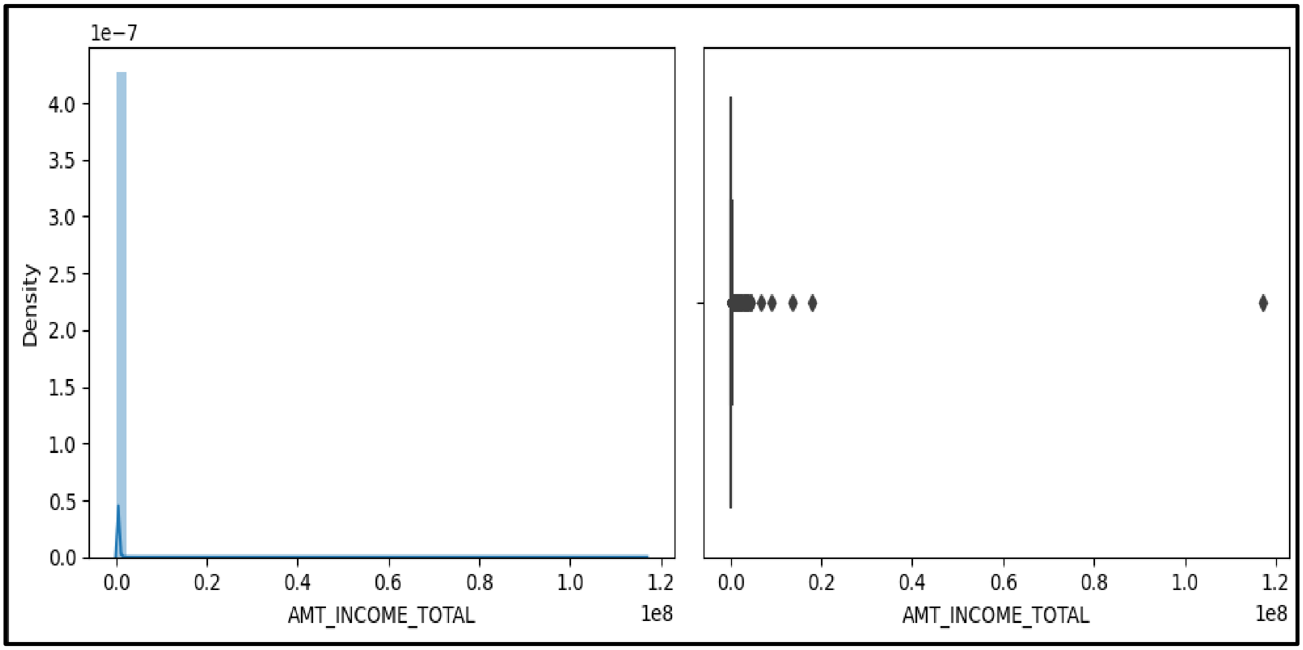
1. 70% of applicants had no children, 19% had 1 child and 8% had 2 children.
2. 51% of applicants had only two family members, 22% had one and 17% had three family members.
3. Most of the applicants were from Region Rating 2.

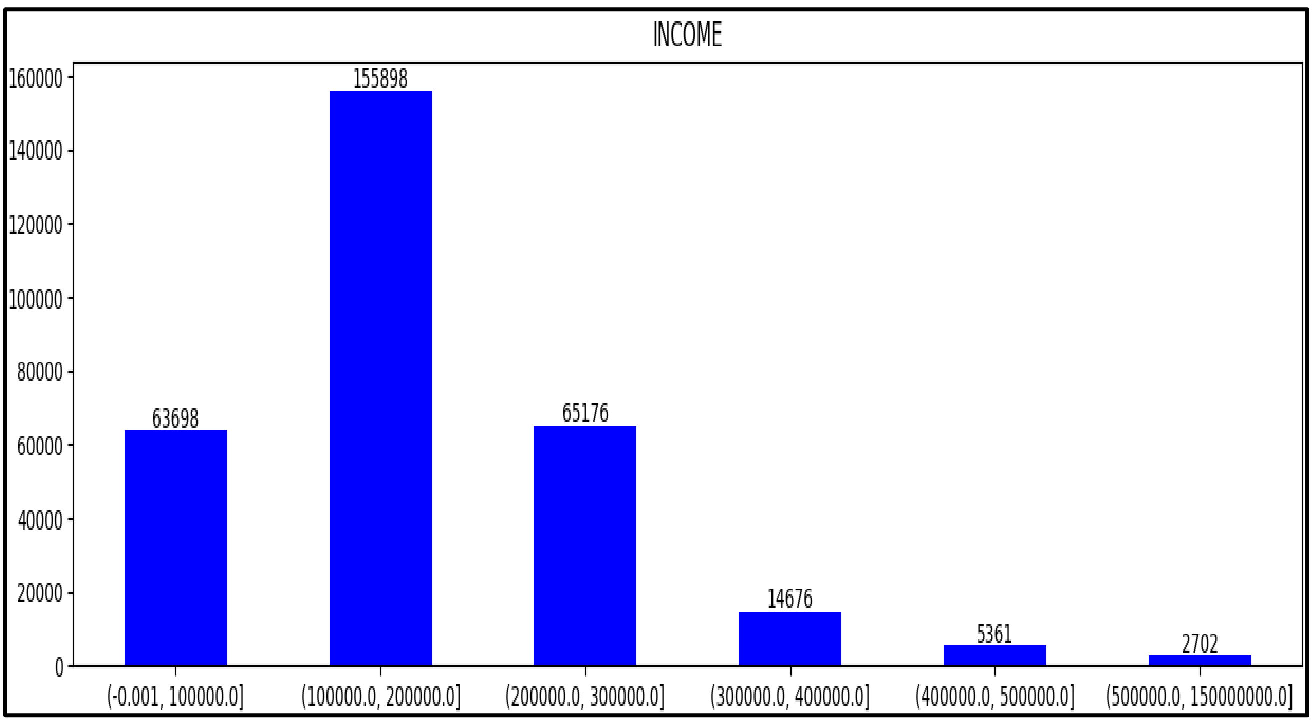




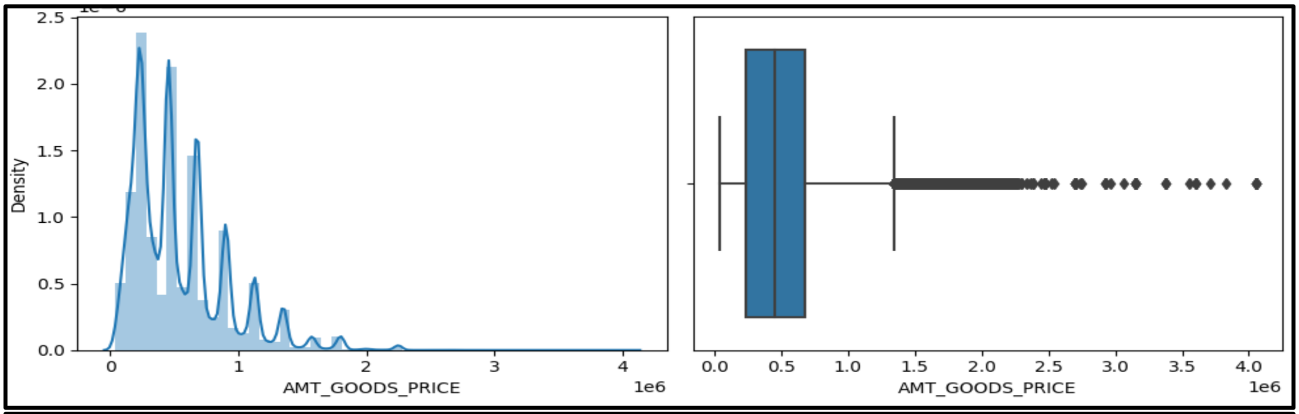
Numerical Columns

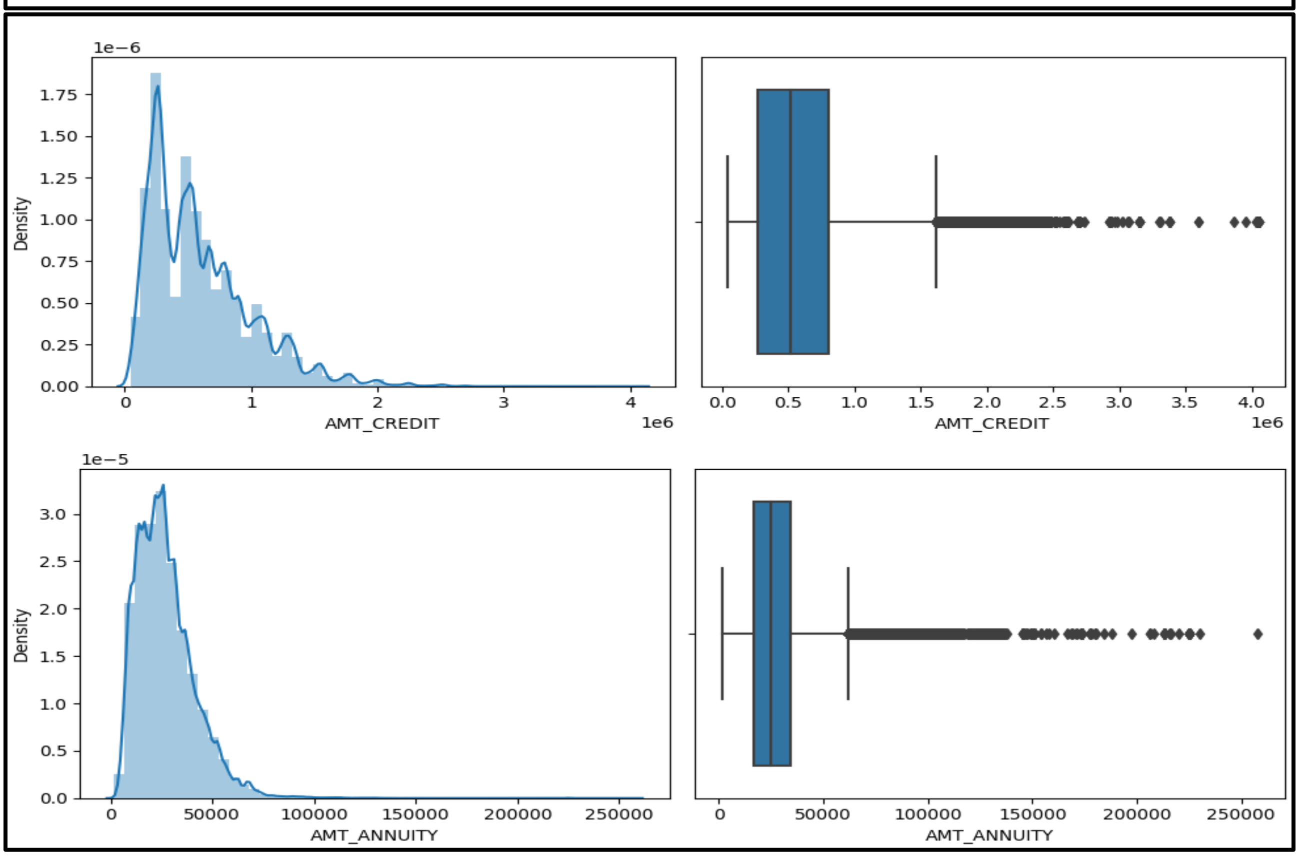
1. 75% of applicants have income up to 2 Lac. The income range with highest number of applicants was between 1Lac and 2 Lac.
2. The minimum income of an applicant is 25,000 while maximum is 11.7 crore. However, 99% of applicants have income below 5 Lac.



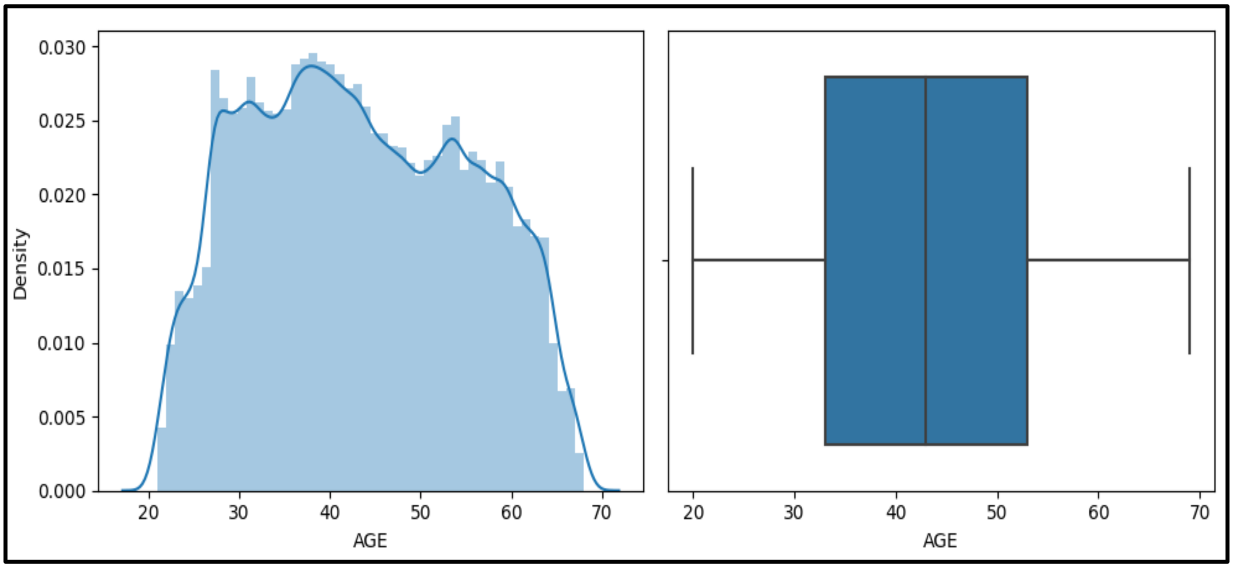


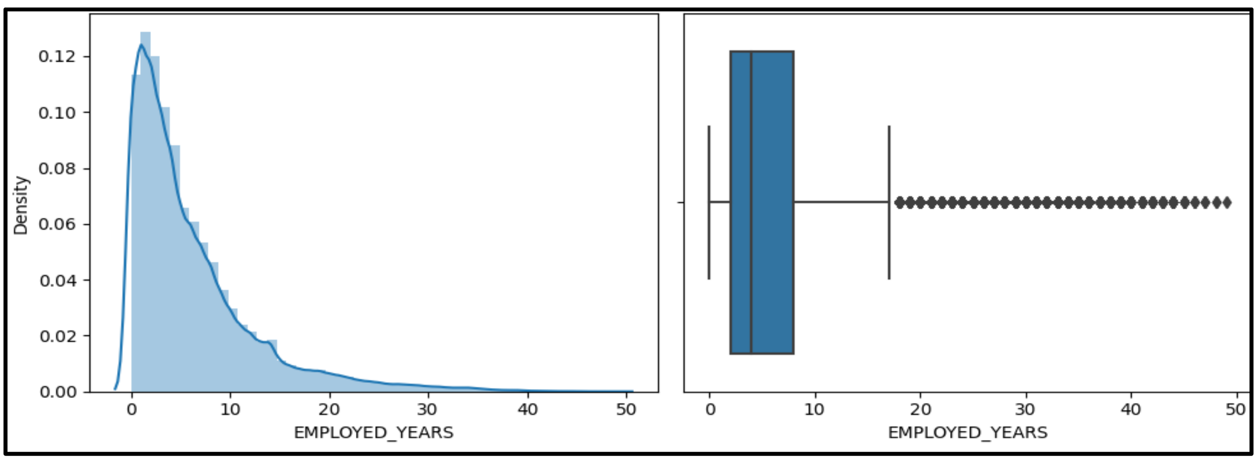
3. The goods price range with maximum number of applicants was from 2 to 4 Lac. 75% of the applicants filed for loan against a goods' price value under 6.7 Lac. The minimum Goods price was about 40K and maximum was 40 Lac.
4. The range of credit approved for maximum number of applicants was from 2 to 4 Lac. 75% of the applications had Credit approved till the amount of 8 Lac. The maximum amount approved was of 40 Lac.
5. 75% of the applicants paid an Annuity amount below 35K. The maximum Annuity amount was of 2.6 Lac.





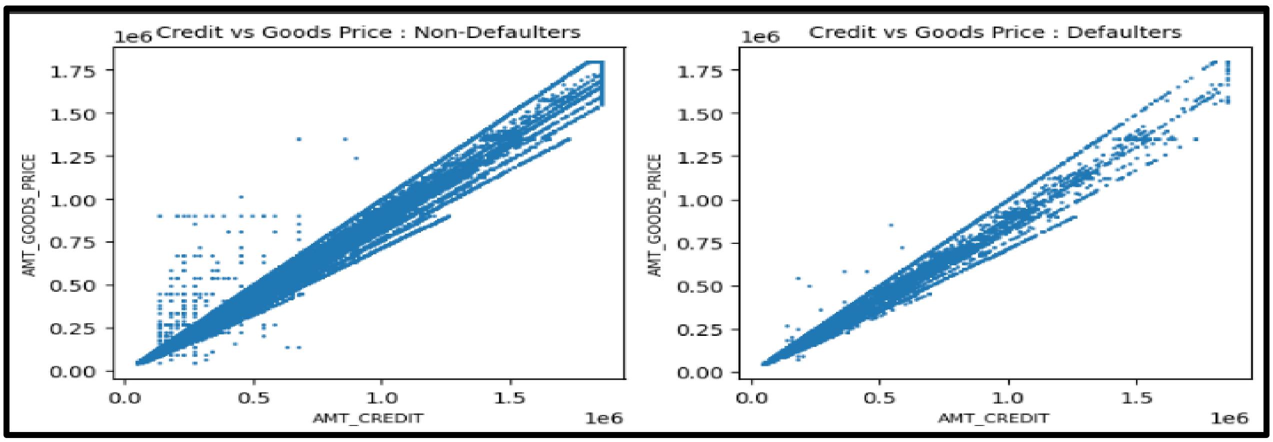
6. Most of the applicants were in Age range 33 to 53.
7. Most of the applicants were employed for 2 to 8 years.
8. Most of the applicants had changed their registration in last 5 to 20 years.



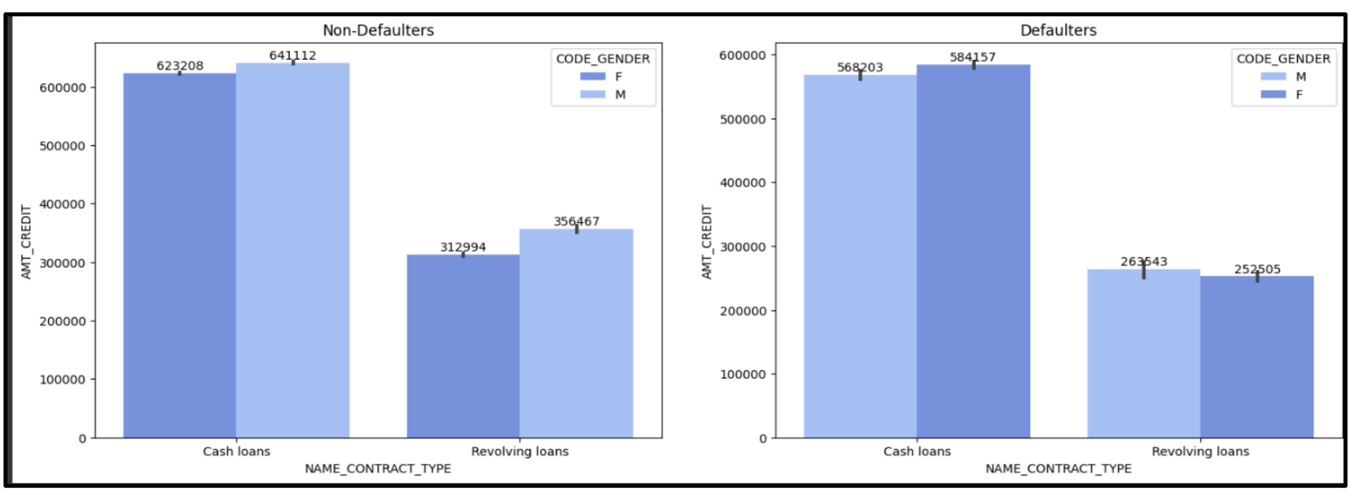


Bivariate Analysis

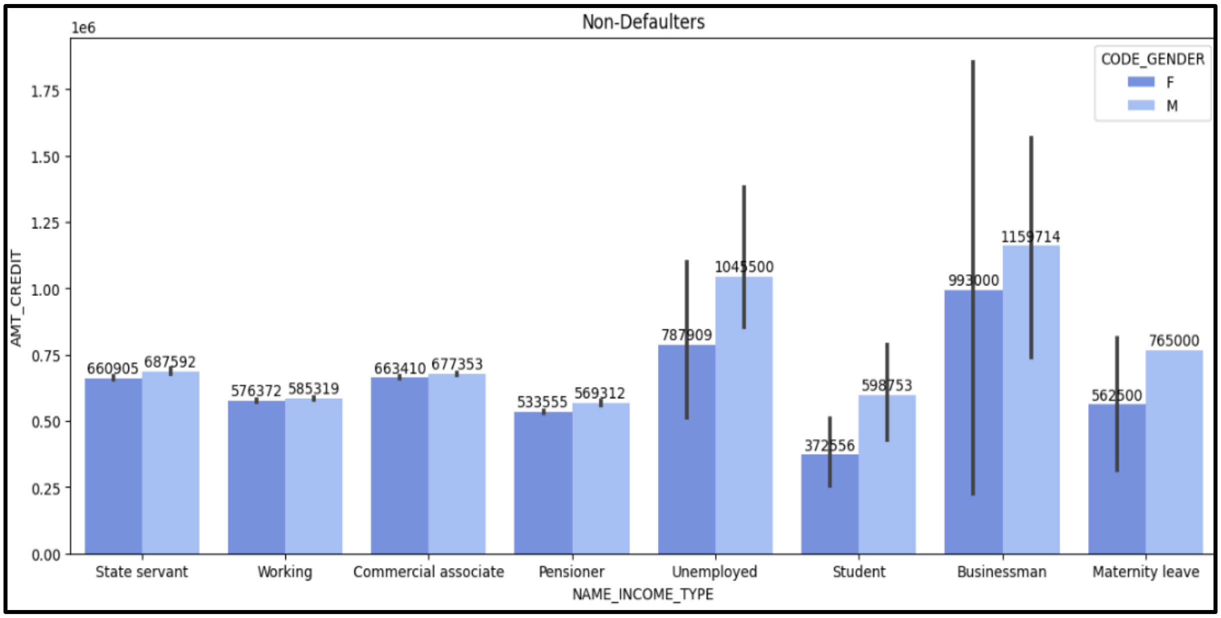
1. Higher correlation between features OBS_60_CNT_SOCIAL_CIRCLES and OBS_30_CNT_SOCIAL_CIRCLES was observed.
2. Similarly, higher correlation between features DEF_60_CNT_SOCIAL_CIRCLES and DEF_30_CNT_SOCIAL_CIRCLES was observed.
3. There is high correlation between goods' price and credit amount for both defaulters and non-defaulters.
4. It was observed that with an increase in income, there was an increase in credit amount.

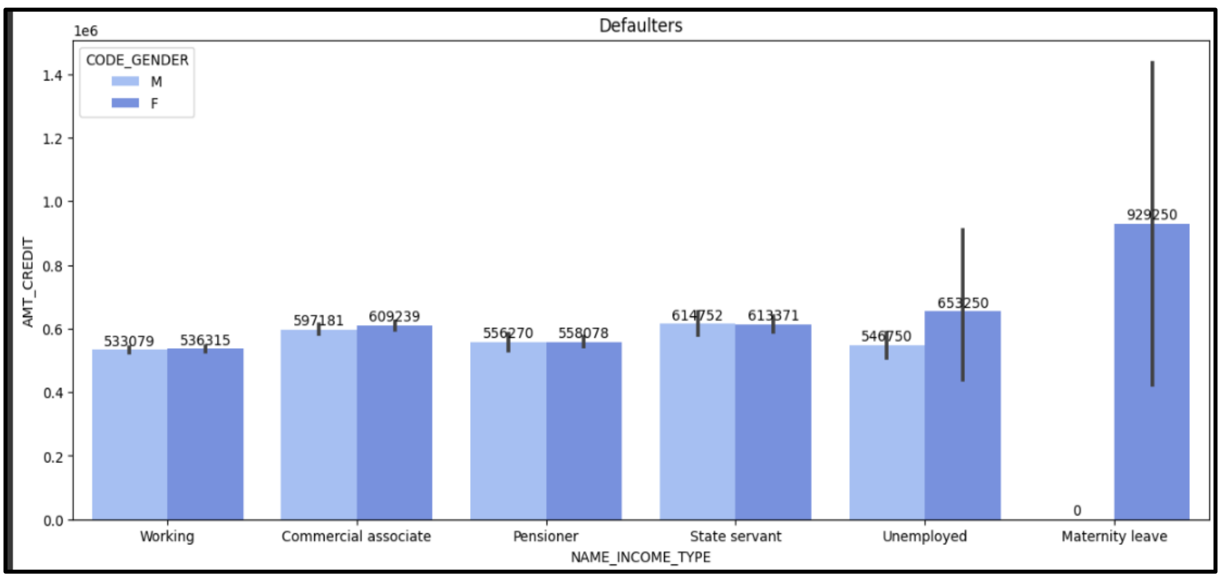


5. There was no correlation between EXT_SOURCE_2 and EXT_SOURCE_3.
6. Credit amount was higher for Cash loans. Moreover, for non-defaulters, the number of male applicants was higher for both cash as well as revolving loans.



7. The credit amount was highest for the male applicants with an academic degree, followed by male applicants with higher education.
8. Amongst non-defaulters, male businessmen and male unemployed had the highest credit amount. For defaulters, female on maternity leave or unemployed had the highest credit amount.





9. For non-defaulters, males earned more irrespective of profession, with an exception of business and student income. For defaulters, the male applicants earned more, with an exception of unemployed and maternity leave.

Segmented Univariate Analysis

Categorical Columns

The defaulter percentage for each value of each categorical column is shown in the clips below.

NAME_CONTRACT_TYPE

	Value	Default_Percentage
0	Cash loans	8.345913
1	Revolving loans	5.478329

CODE_GENDER

	Value	Default_Percentage
0	M	10.14192
1	F	6.99919

NAME_EDUCATION_TYPE

	Value	Default_Percentage
3	Lower secondary	10.927673
0	Secondary / secondary special	8.939929
2	Incomplete higher	8.484966
1	Higher education	5.355115
4	Academic degree	1.829268

NAME_TYPE_SUITE

	Value	Default_Percentage
6	Other_B	9.830508
4	Other_A	8.775982
7	Group of people	8.487085
0	Unaccompanied	8.183047
2	Spouse, partner	7.871592
1	Family	7.494583
3	Children	7.376798
5	NaN	0.000000

NAME_FAMILY_STATUS

	Value	Default_Percentage
2	Civil marriage	9.944584
0	Single / not married	9.807675
4	Separated	8.194234
1	Married	7.559791
3	Widow	5.824217

NAME_HOUSING_TYPE

	Value	Default_Percentage
1	Rented apartment	12.313051
2	With parents	11.698113
3	Municipal apartment	8.539748
5	Co-op apartment	7.932264
0	House / apartment	7.795711
4	Office apartment	6.572411

FLAG_OWN_CAR

	Value	Default_Percentage
0	N	8.500227
1	Y	7.243730

FLAG_OWN_REALTY

	Value	Default_Percentage
1	N	8.324929
0	Y	7.961577

NAME_INCOME_TYPE

	Value	Default_Percentage
7	Maternity leave	40.000000
4	Unemployed	36.363636
0	Working	9.588472
2	Commercial associate	7.484257
1	State servant	5.754965
3	Pensioner	5.386366
5	Student	0.000000
6	Businessman	0.000000

1. The percentage of defaulters was higher in Cash Loans as compared to revolving loans.
2. Males, while being less in number, defaulted more than women.
3. The applicants with lower secondary education, while less in count, defaulted more than other education types. People with academic degrees defaulted the least.
4. The accommodation type Other_B had the highest percentage of defaulters while people accommodating with family members, especially children, had the smallest default percentage.
5. The applicants with Civil marriage had the most difficulty in repayment, while widows defaulted the least.
6. People living in rented apartments had the highest default percentage while those residing in office apartments had the least difficulty in loan payment.

7. There was negligible difference between people who owned realty/car and people who didn't, with non-owners defaulting more.
8. The people on maternity leave or unemployed had highest default percentage while students and businessmen had no difficulty in payments.

OCCUPATION_TYPE

	Value	Default_Percentage
14	Low-skill Laborers	17.152413
5	Drivers	11.326130
13	Waiters/barmen staff	11.275964
11	Security staff	10.742449
0	Laborers	10.578770
8	Cooking staff	10.443996
6	Sales staff	9.631799
7	Cleaning staff	9.606705
15	Realty agents	7.856192
16	Secretaries	7.049808
10	Medicine staff	6.700246
9	Private service staff	6.598793
17	IT staff	6.463878
18	HR staff	6.394316
1	Core staff	6.303954
3	Managers	6.214028
12	High skill tech staff	6.159930
2	Accountants	4.830327
4	NaN	0.000000

REG_REGION_NOT_LIVE_REGION

	Value	Default_Percentage
1	1.0	9.297831
0	0.0	8.054046

REG_REGION_NOT_WORK_REGION

	Value	Default_Percentage
1	1.0	8.890597
0	0.0	8.029147

LIVE_REGION_NOT_WORK_REGION

	Value	Default_Percentage
1	1.0	8.445973
0	0.0	8.057070

REG_CITY_NOT_LIVE_CITY

	Value	Default_Percentage
1	1.0	12.225966
0	0.0	7.720692

REG_CITY_NOT_WORK_CITY

	Value	Default_Percentage
1	1.0	10.611427
0	0.0	7.312672

LIVE_CITY_NOT_WORK_CITY

	Value	Default_Percentage
1	1.0	9.966495
0	0.0	7.658465

FLAG_MOBIL

	Value	Default_Percentage
0	1.0	8.072908
1	0.0	0.000000

FLAG_EMP_PHONE

	Value	Default_Percentage
0	1.0	8.659990
1	0.0	5.400282

FLAG_WORK_PHONE

	Value	Default_Percentage
1	1.0	9.630065
0	0.0	7.685122

FLAG_CONT_MOBILE

	Value	Default_Percentage
0	1.0	8.073318
1	0.0	7.839721

FLAG_PHONE

	Value	Default_Percentage
1	0.0	8.478379
0	1.0	7.035670

FLAG_EMAIL

	Value	Default_Percentage
0	0.0	8.084628
1	1.0	7.877537

FLAG_DOCUMENT_3

	Value	Default_Percentage
0	1.0	8.844921
1	0.0	6.182503

9. Low skill laborers, drivers, waiters had high default percentage while high skill tech staff and accountants had low default percentage.
10. Surprisingly, the people who had provided mobile numbers, work contact, emails, document 3, etc. defaulted more than the ones who didn't.
11. The people whose contact/work address didn't match permanent address defaulted more than the ones whose did.

Discrete Columns

1. It was observed that defaulter percentage increased with an increase in the count of children/family members.
2. Also the Region rating 3 had highest default percentage, followed by Region rating 2. Region rating 1 had the least default percentage.
3. As the observations of client's social surroundings with defaults increased, the default percentage also increased.
4. The clients with higher number of enquiries to Credit Bureau in last one year (excluding last 3 months before application) had higher default percentage.

AMT_REQ_CREDIT_BUREAU_YEAR

	Value	Default_Percentage
7	8.0	9.363853
9	7.0	9.201344
8	6.0	9.071336
5	5.0	8.322270
4	4.0	8.255286
3	2.0	8.104877
6	3.0	7.957654
0	1.0	7.333806
1	0.0	7.134998
2	NaN	0.000000

CNT_CHILDREN

	Value	Default_Percentage
3	3.0	10.042135
1	1.0	8.923575
2	2.0	8.721821
0	0.0	7.711809

CNT_FAM_MEMBERS

	Value	Default_Percentage
4	5.0	9.907662
2	3.0	8.760290
3	4.0	8.648824
0	1.0	8.364408
1	2.0	7.583498
5	NaN	0.000000

DEF_60_CNT_SOCIAL_CIRCLE

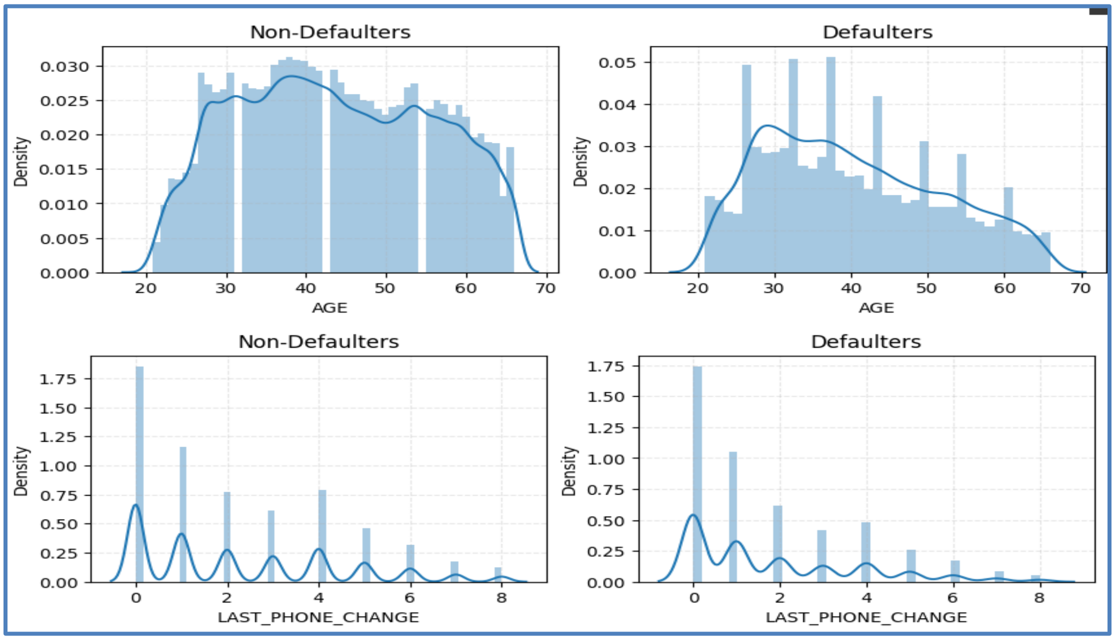
	Value	Default_Percentage
0	2.0	12.678208
2	1.0	10.516918
1	0.0	7.834825
3	NaN	0.000000

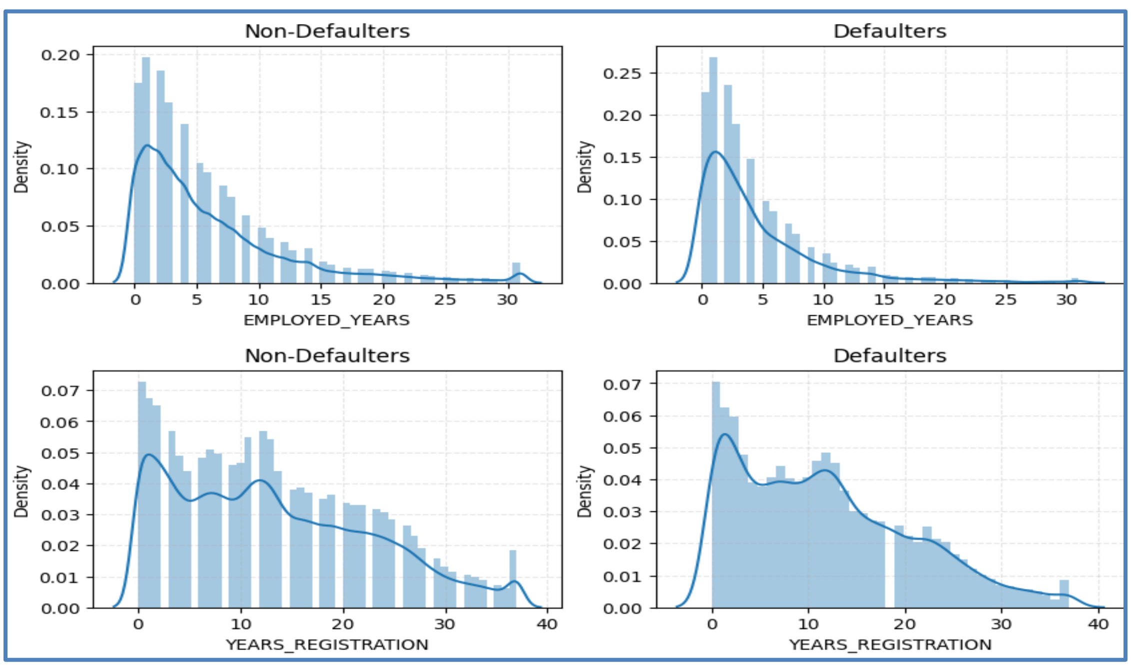
REGION_RATING_CLIENT

	Value	Default_Percentage
2	3.0	11.102835
0	2.0	7.889102
1	1.0	4.820325

Numerical Columns

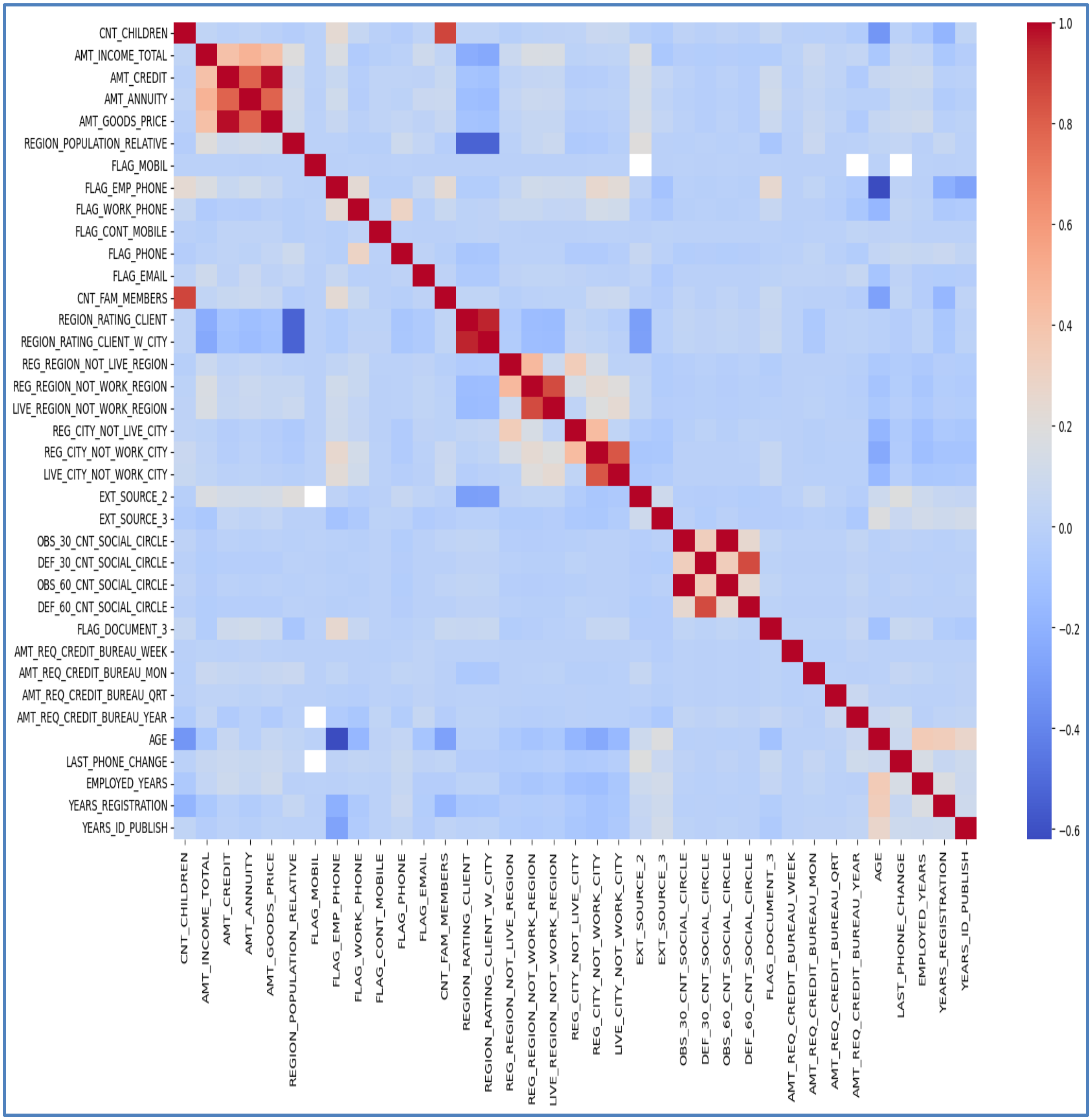
1. The people with age around 28-30 years defaulted the most.
2. The people who had changed their phone number less than a year before also defaulted the most.
3. The people who had been employed for less than 5 years defaulted more than others.
4. The people who had registered less than 5 years before defaulted the most.





Correlation

application_data



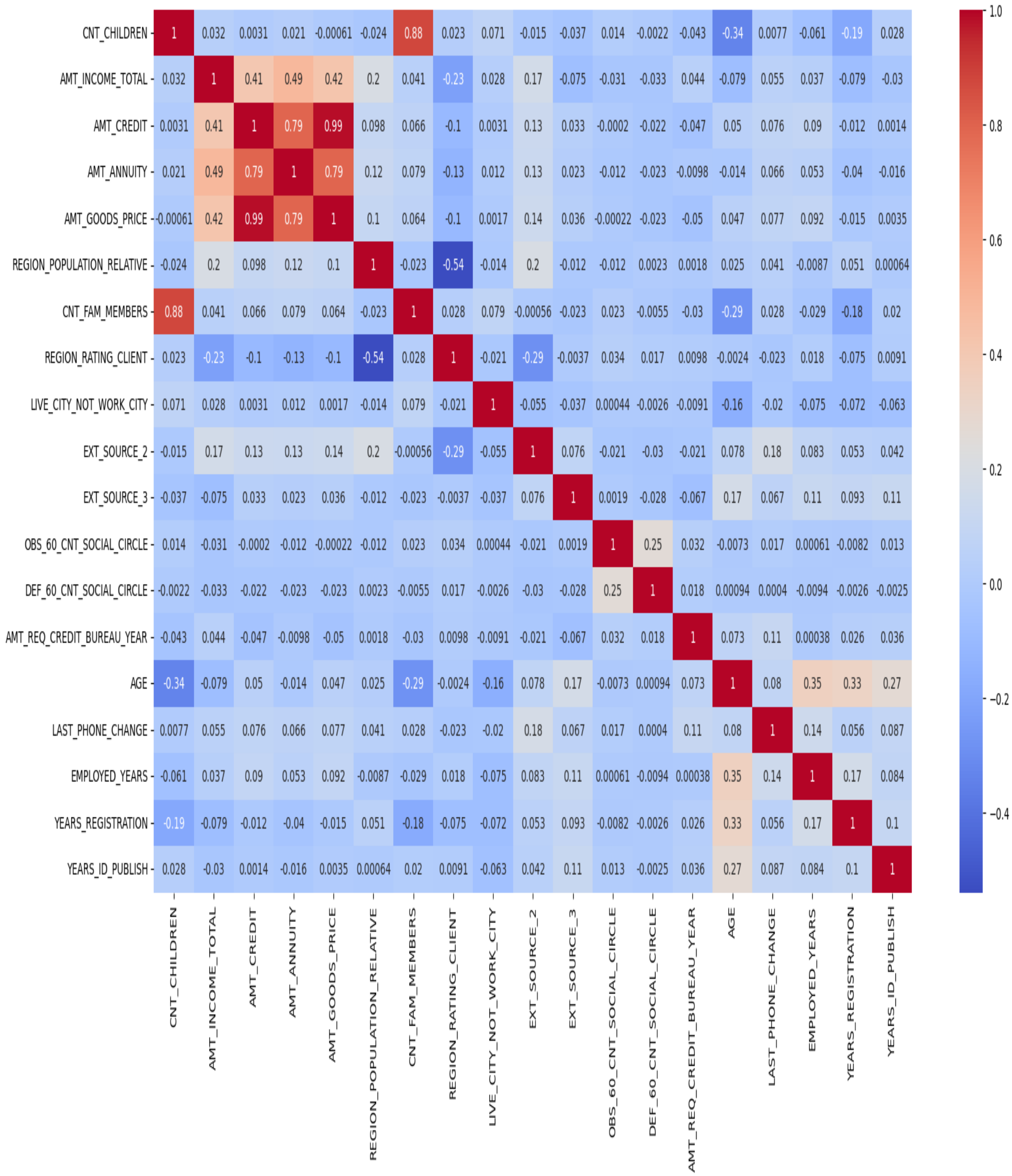
Top 10 correlation for application_data

```
[ ] # Top 10 correlation for dataframe df1
    corr_sorted.tail(20)
```

```
AGE                          FLAG_EMP_PHONE                 0.619204
FLAG_EMP_PHONE               AGE                            0.619204
AMT_ANNUITY                  AMT_CREDIT                     0.787751
AMT_CREDIT                   AMT_ANNUITY                    0.787751
AMT_GOODS_PRICE              AMT_ANNUITY                    0.790507
AMT_ANNUITY                  AMT_GOODS_PRICE                0.790507
LIVE_CITY_NOT_WORK_CITY      REG_CITY_NOT_WORK_CITY         0.825575
REG_CITY_NOT_WORK_CITY       LIVE_CITY_NOT_WORK_CITY        0.825575
DEF_30_CNT_SOCIAL_CIRCLE     DEF_60_CNT_SOCIAL_CIRCLE       0.860517
DEF_60_CNT_SOCIAL_CIRCLE     DEF_30_CNT_SOCIAL_CIRCLE       0.860517
REG_REGION_NOT_WORK_REGION   LIVE_REGION_NOT_WORK_REGION    0.860627
LIVE_REGION_NOT_WORK_REGION  REG_REGION_NOT_WORK_REGION     0.860627
CNT_FAM_MEMBERS              CNT_CHILDREN                   0.879161
CNT_CHILDREN                 CNT_FAM_MEMBERS                0.879161
REGION_RATING_CLIENT_W_CITY  REGION_RATING_CLIENT           0.950842
REGION_RATING_CLIENT         REGION_RATING_CLIENT_W_CITY    0.950842
AMT_CREDIT                   AMT_GOODS_PRICE                0.986432
AMT_GOODS_PRICE              AMT_CREDIT                     0.986432
OBS_60_CNT_SOCIAL_CIRCLE     OBS_30_CNT_SOCIAL_CIRCLE       0.998490
OBS_30_CNT_SOCIAL_CIRCLE     OBS_60_CNT_SOCIAL_CIRCLE       0.998490
dtype: float64
```

<u>Non-defaulter data</u>

Heatmap for Target = 0



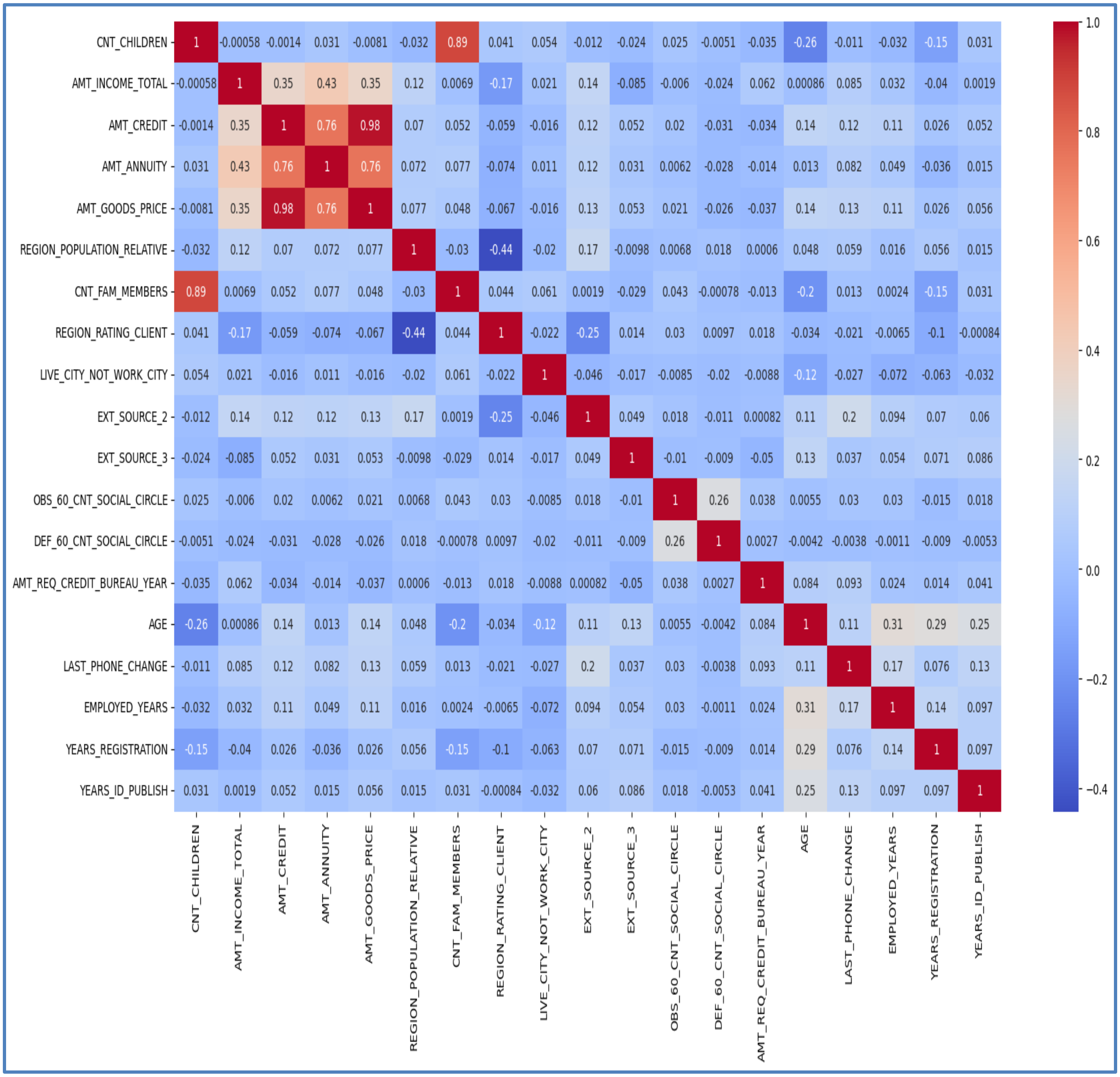
Top 10 Correlation for Non-defaulters

CNT_CHILDREN	AGE	0.336992
AGE	CNT_CHILDREN	0.336992
	EMPLOYED_YEARS	0.350076
EMPLOYED_YEARS	AGE	0.350076
AMT_CREDIT	AMT_INCOME_TOTAL	0.410460
AMT_INCOME_TOTAL	AMT_CREDIT	0.410460
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.417296
AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.417296
	AMT_ANNUITY	0.488409
AMT_ANNUITY	AMT_INCOME_TOTAL	0.488409
REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT	0.539005
REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	0.539005
AMT_ANNUITY	AMT_CREDIT	0.789835
AMT_CREDIT	AMT_ANNUITY	0.789835
AMT_GOODS_PRICE	AMT_ANNUITY	0.792956
AMT_ANNUITY	AMT_GOODS_PRICE	0.792956
CNT_CHILDREN	CNT_FAM_MEMBERS	0.878571
CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
AMT_GOODS_PRICE	AMT_CREDIT	0.986732
AMT_CREDIT	AMT_GOODS_PRICE	0.986732

dtype: float64

<u>Defaulter data</u>

Heatmap



Top 10 correlations for Defaulters

```
YEARS_REGISTRATION          AGE                          0.287475
AGE                         YEARS_REGISTRATION           0.287475
EMPLOYED_YEARS              AGE                          0.305951
AGE                         EMPLOYED_YEARS               0.305951
AMT_INCOME_TOTAL            AMT_CREDIT                   0.350124
AMT_CREDIT                  AMT_INCOME_TOTAL             0.350124
AMT_GOODS_PRICE             AMT_INCOME_TOTAL             0.352770
AMT_INCOME_TOTAL            AMT_GOODS_PRICE              0.352770
AMT_ANNUITY                 AMT_INCOME_TOTAL             0.427960
AMT_INCOME_TOTAL            AMT_ANNUITY                  0.427960
REGION_POPULATION_RELATIVE  REGION_RATING_CLIENT         0.443236
REGION_RATING_CLIENT        REGION_POPULATION_RELATIVE   0.443236
AMT_ANNUITY                 AMT_GOODS_PRICE              0.757730
AMT_GOODS_PRICE             AMT_ANNUITY                  0.757730
AMT_CREDIT                  AMT_ANNUITY                  0.758001
AMT_ANNUITY                 AMT_CREDIT                   0.758001
CNT_FAM_MEMBERS             CNT_CHILDREN                 0.885484
CNT_CHILDREN                CNT_FAM_MEMBERS              0.885484
AMT_GOODS_PRICE             AMT_CREDIT                   0.982440
AMT_CREDIT                  AMT_GOODS_PRICE              0.982440
dtype: float64
```

Working with pervious application data

Description

The dataframe prev_app_data has 37 columns and 1670214 rows. There are 15 columns with float datatype, 6 with integer and 16 with object datatype.

Columns Names:

```
['SK_ID_PREV',
 'SK_ID_CURR',
 'NAME_CONTRACT_TYPE',
 'AMT_ANNUITY',
 'AMT_APPLICATION',
 'AMT_CREDIT',
 'AMT_DOWN_PAYMENT',
 'AMT_GOODS_PRICE',
 'WEEKDAY_APPR_PROCESS_START',
 'HOUR_APPR_PROCESS_START',
 'FLAG_LAST_APPL_PER_CONTRACT',
 'NFLAG_LAST_APPL_IN_DAY',
 'RATE_DOWN_PAYMENT',
 'RATE_INTEREST_PRIMARY',
 'RATE_INTEREST_PRIVILEGED',
 'NAME_CASH_LOAN_PURPOSE',
 'NAME_CONTRACT_STATUS',
 'DAYS_DECISION',
 'NAME_PAYMENT_TYPE',
```

```
 'NAME_PAYMENT_TYPE',
 'CODE_REJECT_REASON',
 'NAME_TYPE_SUITE',
 'NAME_CLIENT_TYPE',
 'NAME_GOODS_CATEGORY',
 'NAME_PORTFOLIO',
 'NAME_PRODUCT_TYPE',
 'CHANNEL_TYPE',
 'SELLERPLACE_AREA',
 'NAME_SELLER_INDUSTRY',
 'CNT_PAYMENT',
 'NAME_YIELD_GROUP',
 'PRODUCT_COMBINATION',
 'DAYS_FIRST_DRAWING',
 'DAYS_FIRST_DUE',
 'DAYS_LAST_DUE_1ST_VERSION',
 'DAYS_LAST_DUE',
 'DAYS_TERMINATION',
 'NFLAG_INSURED_ON_APPROVAL']
```

Irrelevant Columns

The following columns with Null values > 50% were removed and the data was stored in df2.

```
AMT_DOWN_PAYMENT              53.636480
RATE_DOWN_PAYMENT             53.636480
RATE_INTEREST_PRIMARY         99.643698
RATE_INTEREST_PRIVILEGED      99.643698
dtype: float64
```

Other irrelevant columns that were removed were:

```
'SK_ID_PREV', 'WEEKDAY_APPR_PROCESS_START', 'SELLERPLACE_AREA',
'HOUR_APPR_PROCESS_START'
```

The data frame df2 has 1670214 rows and 29 columns.

Duplicates

There were 74871 duplicate rows in df2.

After removal of these rows, df2 has 1595343 rows and 29 columns.

Univariate Analysis

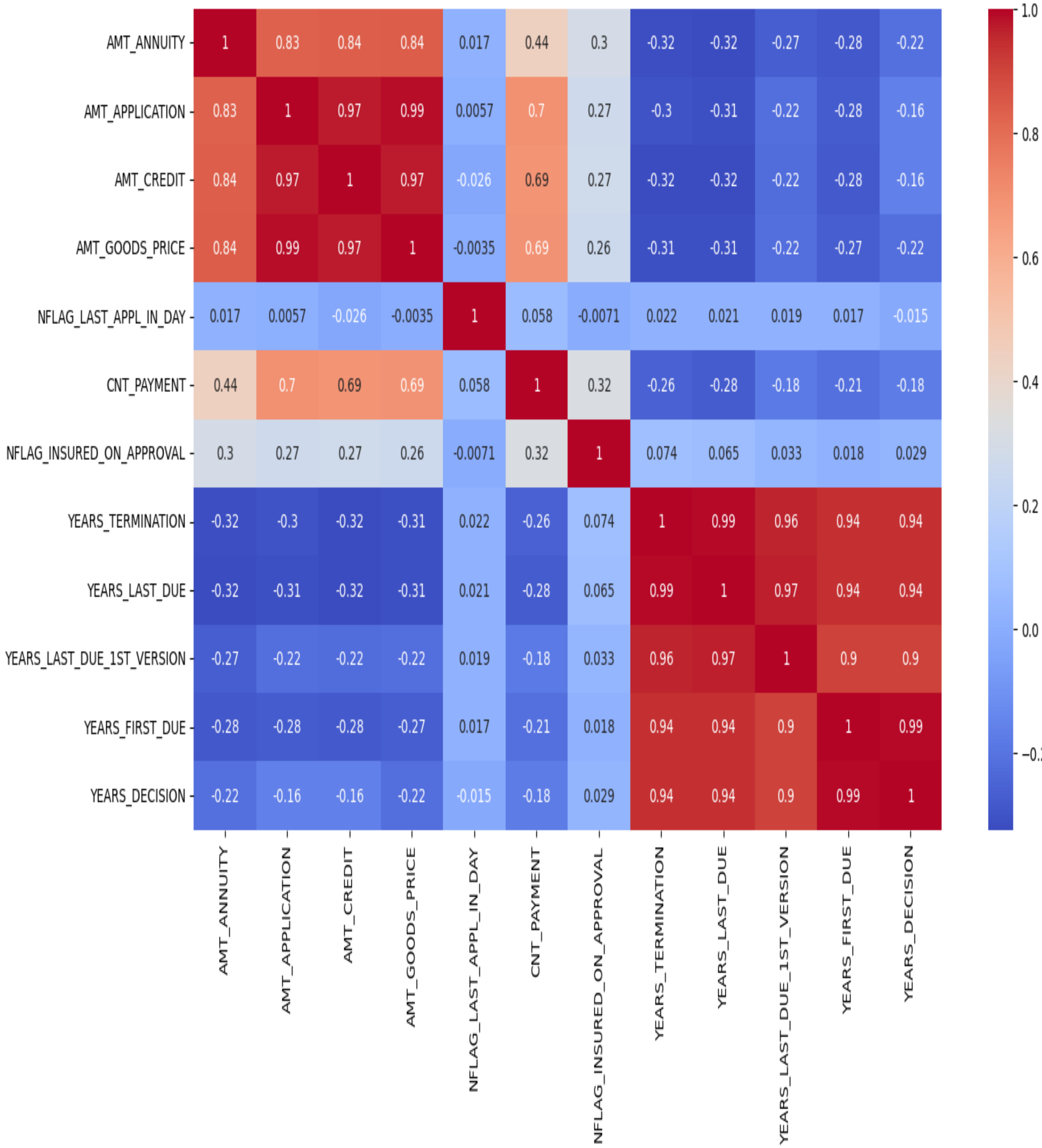
Categorical Columns

1. Among previous applications, 45% of the application were for consumer loan, 42% for cash loan and about 11% for revolving loans.
2. 65% of the previous applications were approved, 17% were refused, and about 15% were canceled while 1.5% of the offers went unused.
3. At least 64% of the applicants made cash payments through banks.
4. Amongst previous applications, 72% of the applicants were older clients, 18% were new applicants while 8% were refreshed.
5. 43% of applications were made for POS, 28% for cash and about 9% for Cards. Less than 0.1% of applications were made for Cars.
6. 24% of the applications had medium interest rate, 22% had high interest rate, 20% had low normal rate while 5% had low action rate.

Numerical Columns

1. 75% of the applications were made for a loan amount less than 2lac. The maximum amount for which application was made was of 69 Lac.
2. Most of the applications had credit amount approved up to 2.25 Lac. The maximum credit amount approved was of 69 Lac.
3. Most of the applicants paid an annuity amount between 7000 and 17,000. The maximum annuity amount paid was of 4 Lac. The highest count for an annuity amount was for 10k.
4. For most of the applicants, their last application had terminated 0 to 4 years before.

Correlation



Merging datasets

The columns in df2 were renamed with 'prev_' as prefix. The data frames df1 and df2 were then merged into merged_df data frame.

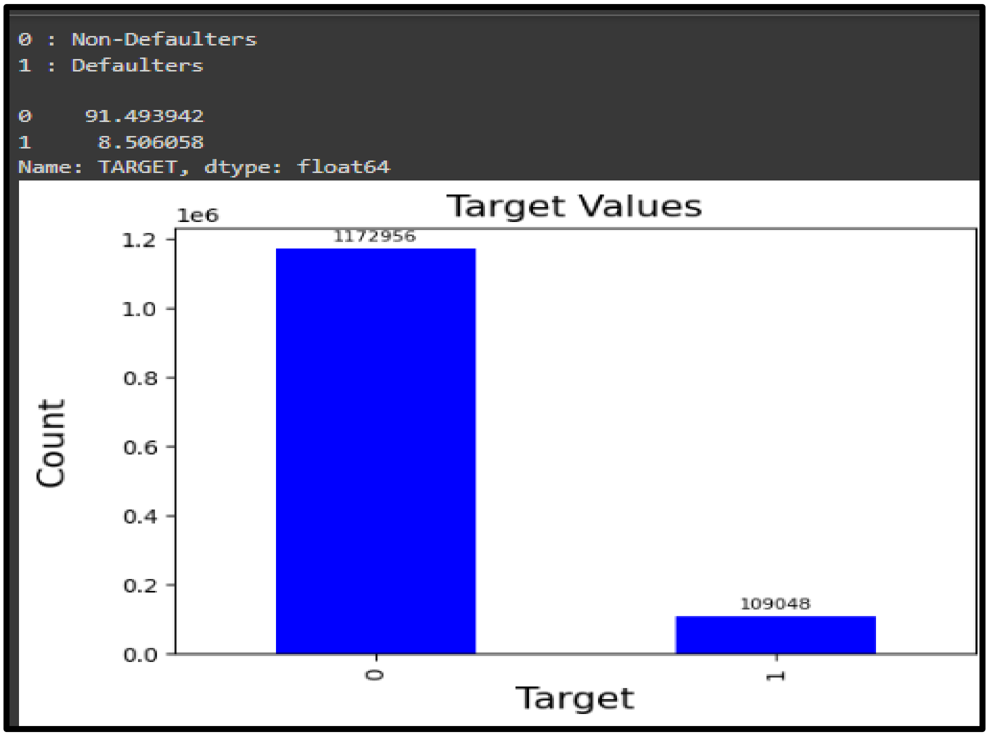
The new data frame merged_df has 1351875 rows and 72 columns. The columns in merged data frame are:

```
Index(['SK_ID_CURR', 'TARGET', 'NAME_CONTRACT_TYPE', 'CODE_GENDER',
       'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL',
       'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'NAME_TYPE_SUITE',
       'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS',
       'NAME_HOUSING_TYPE', 'REGION_POPULATION_RELATIVE', 'FLAG_MOBIL',
       'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE',
       'FLAG_EMAIL', 'OCCUPATION_TYPE', 'CNT_FAM_MEMBERS',
       'REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY',
       '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',
       'ORGANIZATION_TYPE', 'EXT_SOURCE_2', 'EXT_SOURCE_3',
       'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE',
       'OBS_60_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE',
       'FLAG_DOCUMENT_3', 'AMT_REQ_CREDIT_BUREAU_WEEK',
       'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_QRT',
       'AMT_REQ_CREDIT_BUREAU_YEAR', 'AGE', 'LAST_PHONE_CHANGE',
       'EMPLOYED_YEARS', 'YEARS_REGISTRATION', 'YEARS_ID_PUBLISH',
       'prev_NAME_CONTRACT_TYPE', 'prev_AMT_ANNUITY', 'prev_AMT_APPLICATION',
       'prev_AMT_CREDIT', 'prev_AMT_GOODS_PRICE',
       'prev_FLAG_LAST_APPL_PER_CONTRACT', 'prev_NFLAG_LAST_APPL_IN_DAY',
       'prev_NAME_CONTRACT_STATUS', 'prev_NAME_PAYMENT_TYPE',
       'prev_NAME_TYPE_SUITE', 'prev_NAME_CLIENT_TYPE', 'prev_NAME_PORTFOLIO',
       'prev_CHANNEL_TYPE', 'prev_CNT_PAYMENT', 'prev_NAME_YIELD_GROUP',
       'prev_PRODUCT_COMBINATION', 'prev_NFLAG_INSURED_ON_APPROVAL',
       'prev_YEARS_TERMINATION', 'prev_YEARS_LAST_DUE',
       'prev_YEARS_LAST_DUE_1ST_VERSION', 'prev_YEARS_FIRST_DUE',
       'prev_YEARS_DECISION'],
      dtype='object')
```

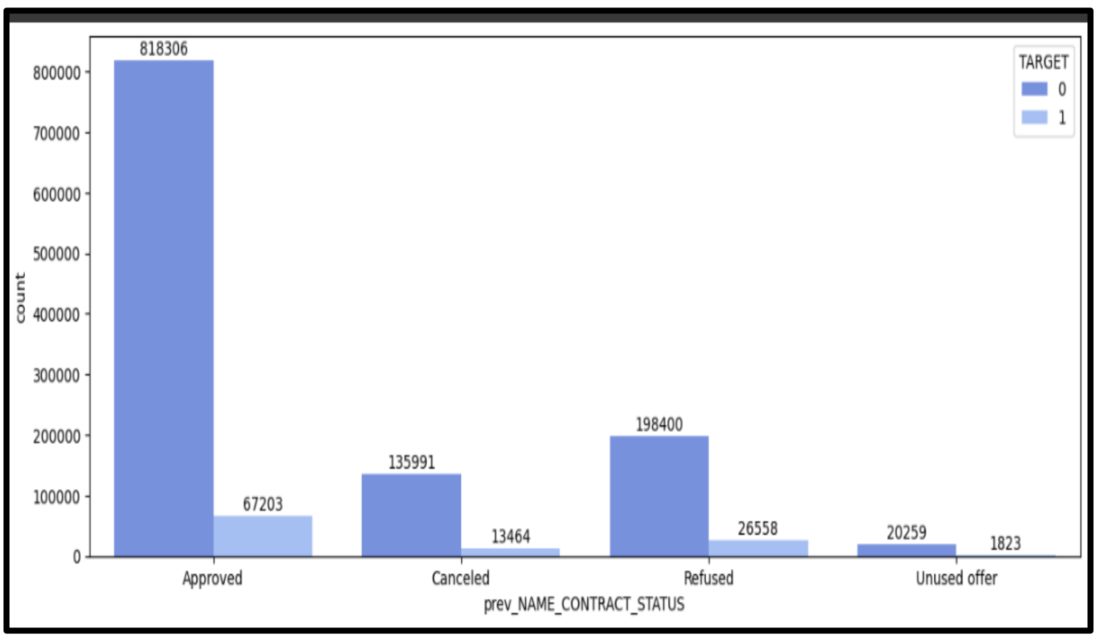
Duplicates

There were 69871 duplicated rows in 72 columns of merged_df. After removal of these rows, there were 1282004 rows left for analysis.

Data Imbalance



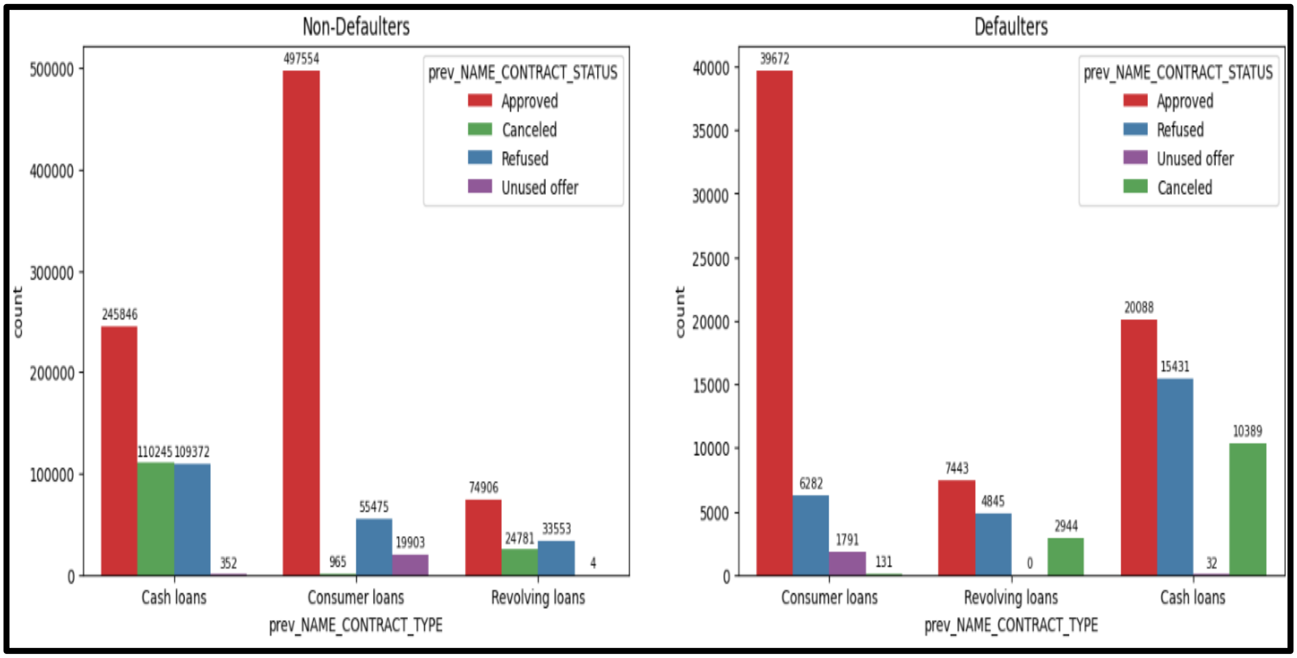
The Non-defaulter to defaulter ratio is 183:17.



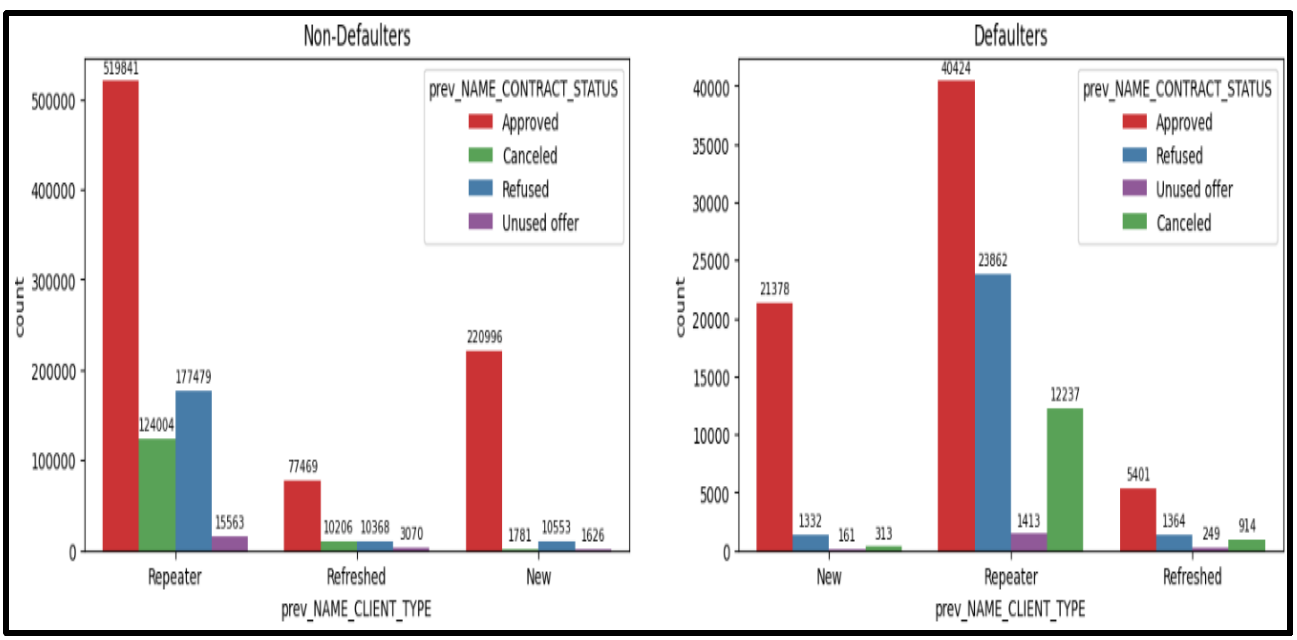
In df2 as well, the amount of approved applications is much higher than other types of applications.

Bivariate Analysis

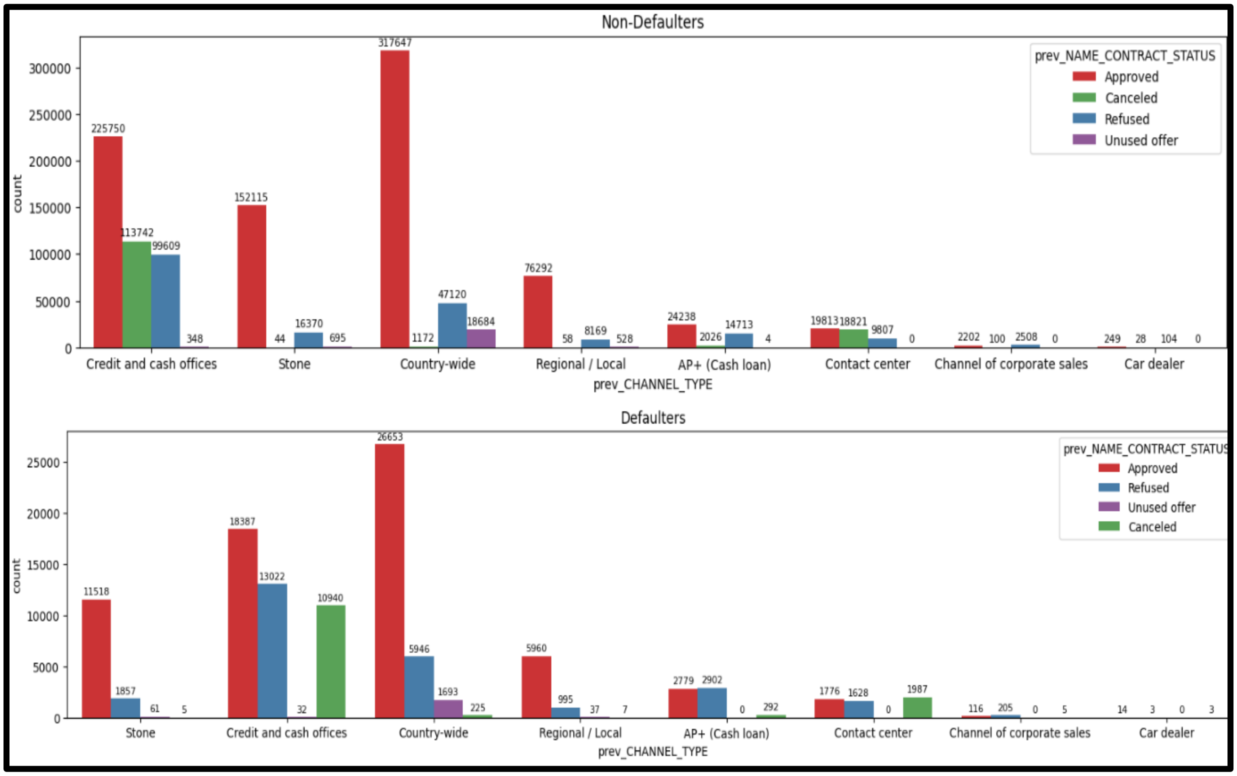
1. The cash loan applications which were refused previously, but approved currently had higher defaulter rate than other applications.



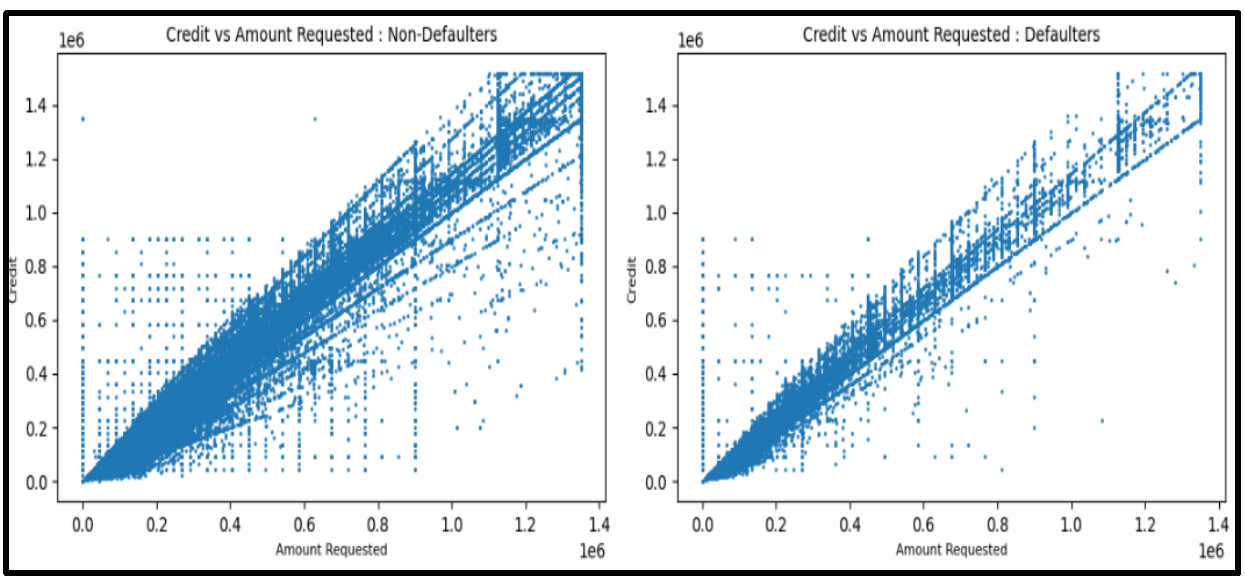
2. Older clients, whose applications had been refused previously, but approved currently, defaulted more than other clients.



3. The number of defaulter clients was higher when acquired through credit and cash offices.



4. Credit approved and amount requested were highly correlated for both defaulters and non-defaulters.



Segmented Univariate for df2

Categorical Columns

The default percentage for each value in Categorical Columns is:

prev_NAME_CONTRACT_TYPE

	Value	Default_Percentage
2	Revolving loans	10.258897
1	Cash loans	8.976952
0	Consumer loans	7.699916

prev_FLAG_LAST_APPL_PER_CONTRACT

	Value	Default_Percentage
1	N	10.835509
0	Y	8.493463

prev_NFLAG_LAST_APPL_IN_DAY

	Value	Default_Percentage
1	0.0	9.952801
0	1.0	8.500538

prev_NAME_CONTRACT_STATUS

	Value	Default_Percentage
2	Refused	11.805759
1	Canceled	9.008732
3	Unused offer	8.255593
0	Approved	7.589194

prev_NAME_PAYMENT_TYPE

	Value	Default_Percentage
2	Non-cash from your account	8.240125
3	Cashless from the account of the employer	8.163265
1	Cash through the bank	8.061016
0	NaN	0.000000

prev_NAME_CLIENT_TYPE

	Value	Default_Percentage
0	New	8.981173
1	Repeater	8.519244
2	Refreshed	7.270660

prev_NAME_PORTFOLIO

	Value	Default_Percentage
3	Cards	10.090615
1	Cash	8.827801
0	POS	7.633301
4	Cars	5.319149
2	NaN	0.000000

prev_CHANNEL_TYPE

	Value	Default_Percentage
4	AP+ (Cash loan)	12.720961
5	Contact center	10.014490
1	Credit and cash offices	8.795841
2	Country-wide	8.235196
3	Regional / Local	7.603807
0	Stone	7.358279
6	Channel of corporate sales	6.347352
7	Car dealer	4.987531

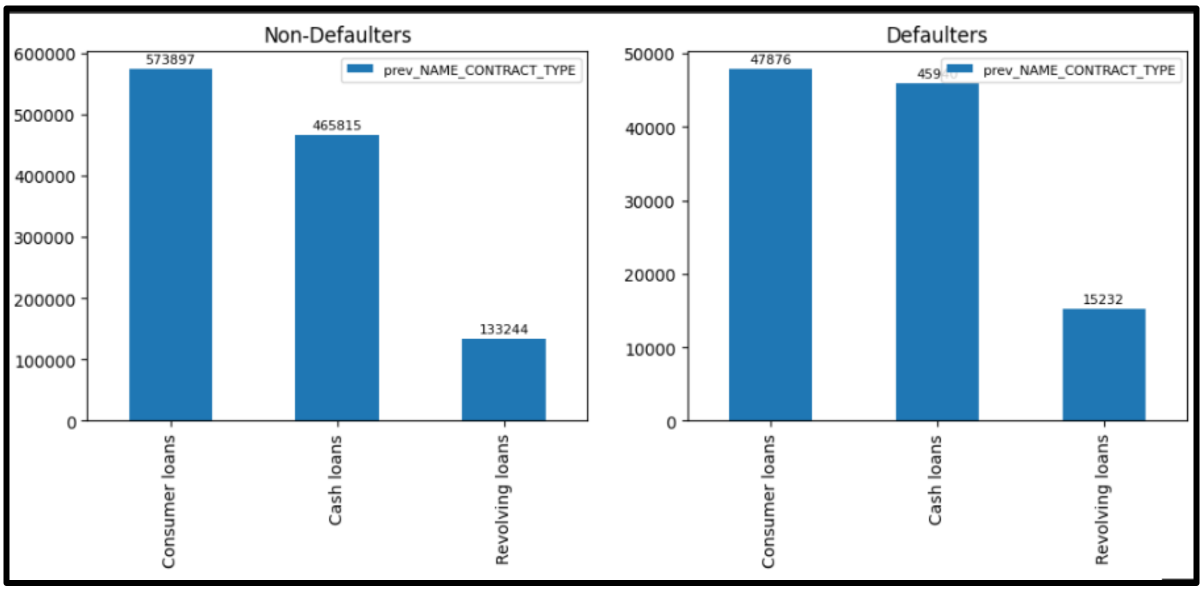
prev_NAME_YIELD_GROUP

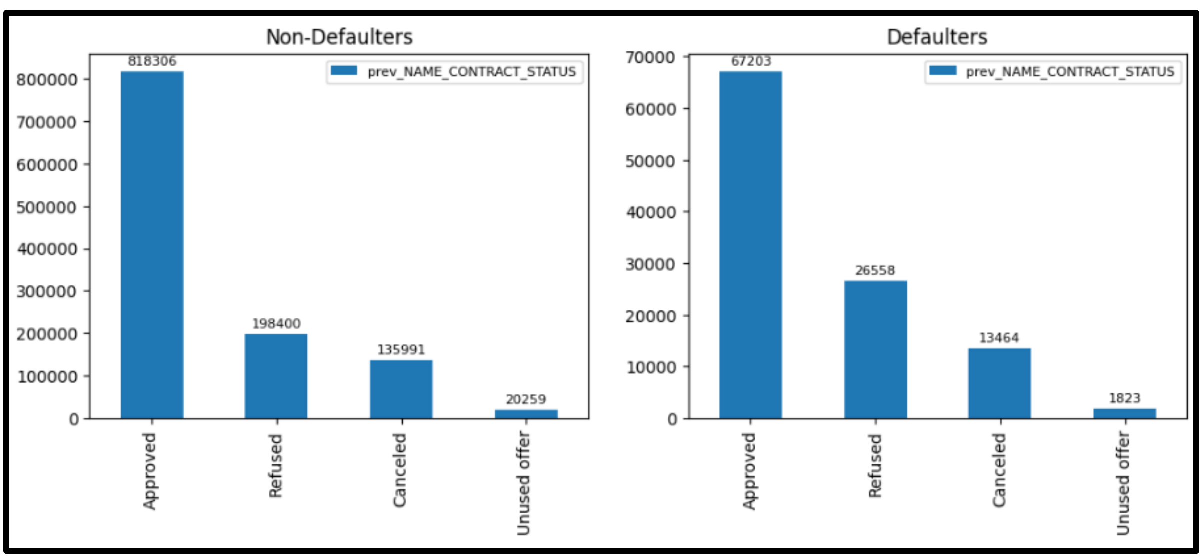
	Value	Default_Percentage
3	high	9.522385
1	middle	8.005877
0	low_normal	7.106717
4	low_action	6.431334
2	NaN	0.000000

prev_PRODUCT_COMBINATION

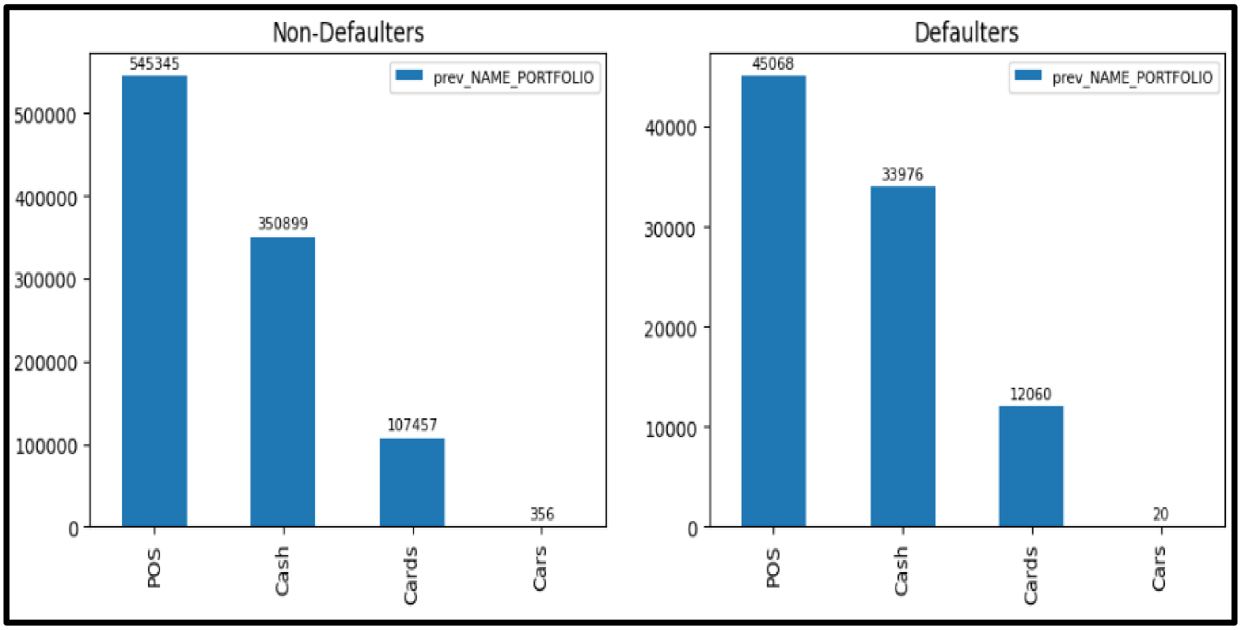
	Value	Default_Percentage
15	Cash Street: middle	11.558956
7	Cash X-Sell: high	11.438637
9	Cash Street: high	11.336487
5	Card Street	11.059885
14	Cash Street: low	10.048957
8	Cash	9.406449
6	Card X-Sell	9.259371
11	POS mobile with interest	8.799329
0	POS other with interest	8.040477
4	POS mobile without interest	7.860968
10	Cash X-Sell: middle	7.787564
3	POS household with interest	7.730250
16	POS others without interest	7.231801
12	POS household without interest	6.652959
1	Cash X-Sell: low	6.531525
2	POS industry with interest	6.284876
13	POS industry without interest	4.650943
17	NaN	0.000000

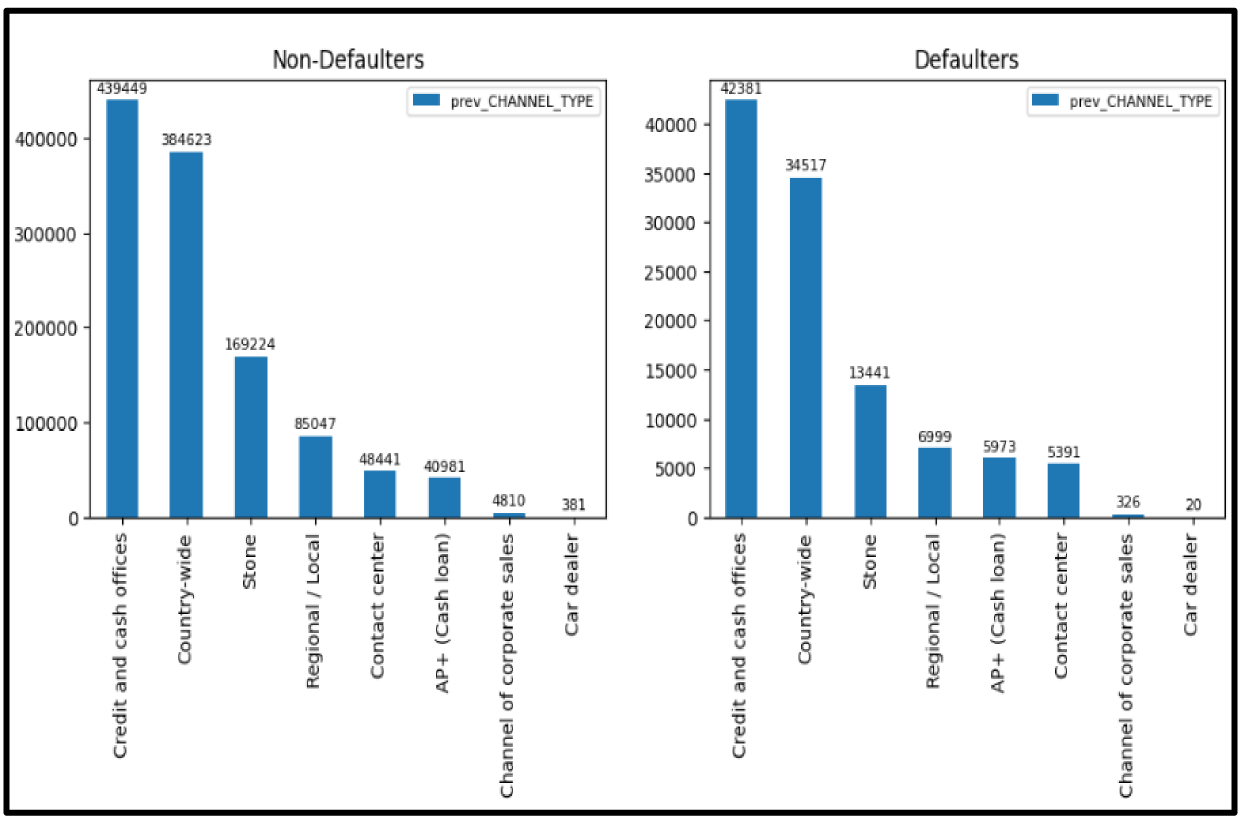
1. Revolving loans have highest default percentage of 10%, followed by cash loans (8.9%) and consumer loans (7.69%).
2. It was observed that the previous applications which were refused (11%), canceled (9%) or went unused (8%) had higher default percentage when approved.
3. The payment type didn't have any significant relationship with change in default percentage.
4. The new clients had a little higher default percentage than the older ones, but nothing significant.



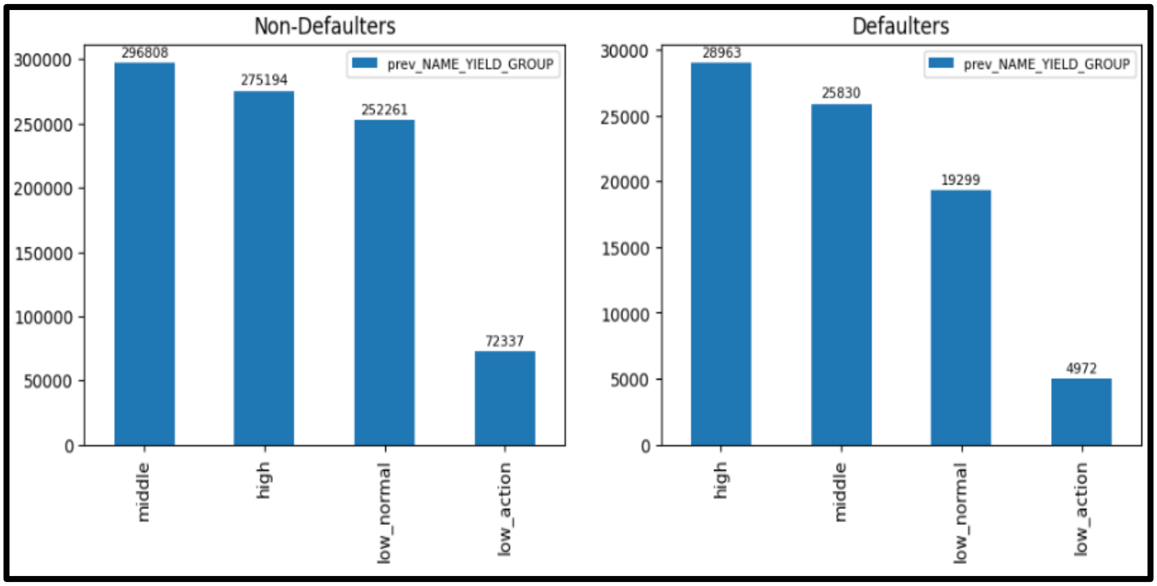


5. The portfolio type Cards had the highest default percentage (10.09%) while Cars had the least (5.3%).
6. The clients acquired through AP+ (Cash loan) and Contact Centre had higher default percentages of 12.7% and 10% respectively.





7. The default percentage increased with the level of yield group. Yield group high had a default percentage of 9.5%.
8. Amongst product combinations, all of Cash Street groups recorded higher default percentage. Cash X-sell high and Card Street also recorded high default percentage.

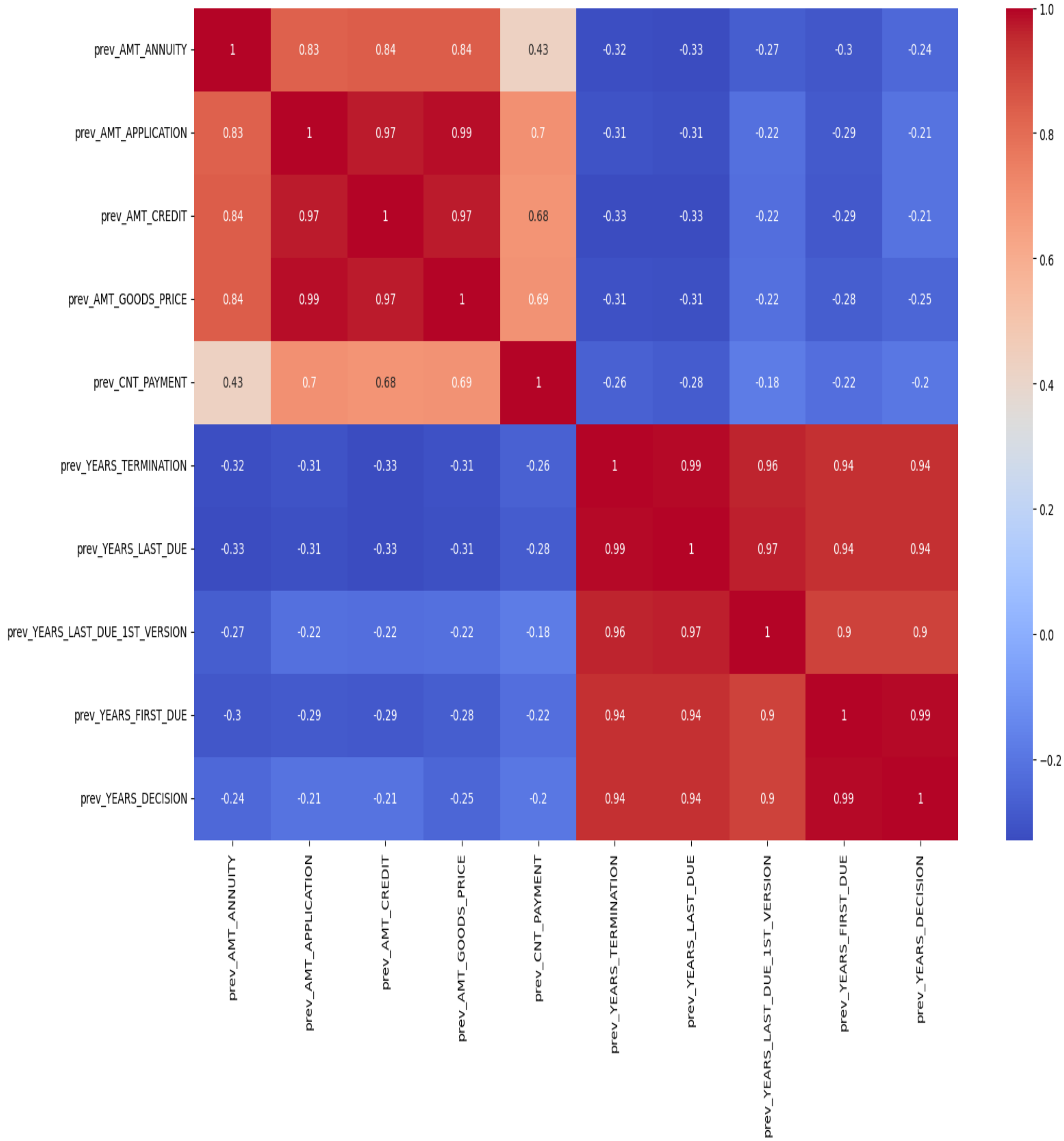


Numerical Columns

For values such as Application Amount, Credit Amount, Annuity Amount, etc., the graphs for defaulters as well as non-defaulters in previous applications followed similar trends.

Correlation

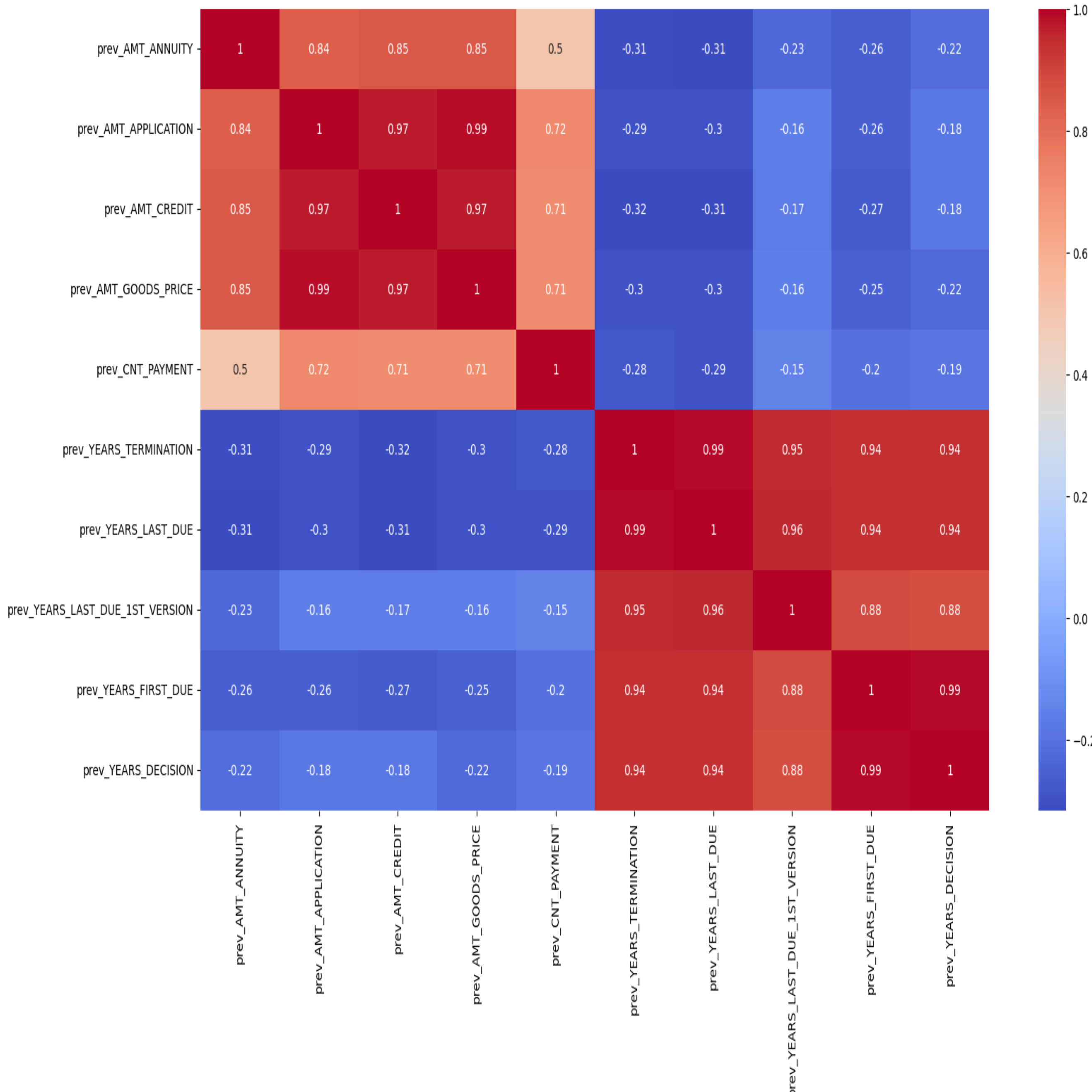
Non-defaulter data



Top 10 Correlations (Non-defaulter)

```
prev_YEARS_FIRST_DUE            prev_YEARS_LAST_DUE             0.940261
prev_YEARS_LAST_DUE             prev_YEARS_FIRST_DUE            0.940261
prev_YEARS_FIRST_DUE            prev_YEARS_TERMINATION          0.940281
prev_YEARS_TERMINATION          prev_YEARS_FIRST_DUE            0.940281
prev_YEARS_LAST_DUE             prev_YEARS_DECISION             0.940392
prev_YEARS_DECISION             prev_YEARS_LAST_DUE             0.940392
prev_YEARS_TERMINATION          prev_YEARS_LAST_DUE_1ST_VERSION 0.962602
prev_YEARS_LAST_DUE_1ST_VERSION prev_YEARS_TERMINATION          0.962602
prev_YEARS_LAST_DUE             prev_YEARS_LAST_DUE_1ST_VERSION 0.966567
prev_YEARS_LAST_DUE_1ST_VERSION prev_YEARS_LAST_DUE             0.966567
prev_AMT_CREDIT                 prev_AMT_APPLICATION            0.970973
prev_AMT_APPLICATION            prev_AMT_CREDIT                 0.970973
prev_AMT_CREDIT                 prev_AMT_GOODS_PRICE            0.971665
prev_AMT_GOODS_PRICE            prev_AMT_CREDIT                 0.971665
prev_AMT_APPLICATION            prev_AMT_GOODS_PRICE            0.989044
prev_AMT_GOODS_PRICE            prev_AMT_APPLICATION            0.989044
prev_YEARS_TERMINATION          prev_YEARS_LAST_DUE             0.990575
prev_YEARS_LAST_DUE             prev_YEARS_TERMINATION          0.990575
prev_YEARS_DECISION             prev_YEARS_FIRST_DUE            0.990926
prev_YEARS_FIRST_DUE            prev_YEARS_DECISION             0.990926
dtype: float64
```

Defaulter data



Top 10 Correlations (Defaulters)

```
prev_YEARS_FIRST_DUE             prev_YEARS_TERMINATION          0.940849
prev_YEARS_TERMINATION           prev_YEARS_FIRST_DUE            0.940849
prev_YEARS_LAST_DUE              prev_YEARS_FIRST_DUE            0.941854
prev_YEARS_FIRST_DUE             prev_YEARS_LAST_DUE             0.941854
prev_YEARS_DECISION              prev_YEARS_LAST_DUE             0.942327
prev_YEARS_LAST_DUE              prev_YEARS_DECISION             0.942327
prev_YEARS_TERMINATION           prev_YEARS_LAST_DUE_1ST_VERSION 0.951892
prev_YEARS_LAST_DUE_1ST_VERSION  prev_YEARS_TERMINATION          0.951892
prev_YEARS_LAST_DUE              prev_YEARS_LAST_DUE_1ST_VERSION 0.956147
prev_YEARS_LAST_DUE_1ST_VERSION  prev_YEARS_LAST_DUE             0.956147
prev_AMT_CREDIT                  prev_AMT_GOODS_PRICE            0.969808
prev_AMT_GOODS_PRICE             prev_AMT_CREDIT                 0.969808
prev_AMT_CREDIT                  prev_AMT_APPLICATION            0.970976
prev_AMT_APPLICATION             prev_AMT_CREDIT                 0.970976
prev_AMT_GOODS_PRICE             prev_AMT_APPLICATION            0.987462
prev_AMT_APPLICATION             prev_AMT_GOODS_PRICE            0.987462
prev_YEARS_DECISION              prev_YEARS_FIRST_DUE            0.990410
prev_YEARS_FIRST_DUE             prev_YEARS_DECISION             0.990410
prev_YEARS_TERMINATION           prev_YEARS_LAST_DUE             0.990606
prev_YEARS_LAST_DUE              prev_YEARS_TERMINATION          0.990606
dtype: float64
```

Final Insights:

1. Males, while being less in number, defaulted more than women.
2. The applicants with lower secondary education and people in low skilled labour defaulted more than other types.
3. The accommodation type Other_B had the higher percentage of defaulters while people accommodating with family members, especially children, had the smaller default percentage. However, it was observed that defaulter percentage increased with an increase in the count of children/family members.
4. People living in rented apartments, on maternity leave or unemployed had the higher default percentage.
5. The people whose contact/work address didn't match permanent address defaulted more than the ones whose did.
6. Region rating 3 had highest default percentage. Moreover, as the observations of client's social surroundings with defaults increased, the default percentage also increased.
7. The clients with higher number of enquiries to Credit Bureau in last one year (excluding last 3 months before application) had higher default percentage.
8. It was observed that the previous applications which were refused (11%), canceled (9%) or went unused (8%) had higher default percentage when approved.
9. The portfolio type Cards had the highest default percentage (10.09%) while Cars had the least (5.3%).
10. The clients acquired through AP+ (Cash loan) and Contact Centre had higher default percentages of 12.7% and 10% respectively.
11. The default percentage increased with the level of yield group. Yield group high had a default percentage of 9.5%.
12. Amongst product combinations, all of Cash Street groups recorded higher default percentage. Cash X-sell high and Card Street also recorded high default percentage.