Credit EDA Assignment

1

Hemavathi AB
ab.hemavathi@gmail.com

Problem Statement 01

This assignment aims to give an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that is learnt in the EDA module, should also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

Business Objective:

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments 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

Data Understanding

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

1. The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
2. All other cases: when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

1. Approved: The Company has approved loan Application
2. Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.
3. Refused: The company had rejected the loan (because the client does not meet their requirements etc.).
4. Unused offer: Loan has been cancelled by the client but at different stagesthe process.

Problem Statement 02

Expectation from Learners:

1. Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.
2. Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)
3. Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.
4. Identify if there is data imbalance in the data. Find the ratio of data imbalance.
5. Analysis for the 'Target variable' in the dataset one can plot in terms of percentage or absolute value to show mix of univariate and bivariate analysis.
6. Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.
7. Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there.

Problem Statement 02

Expectation from Learners:

8. Include visualisations and summarise the most important results in the presentation. Insights should explain why the variable is important for differentiating the clients with payment difficulties with all other cases.
9. You need to submit one/two Ipython notebook which clearly explains the thought process behind your analysis (either in comments of markdown text), code and relevant plots.
9. The presentation file needs to be in PDF format and should contain the points discussed above with the necessary visualisations. Also, all the visualisations and plots must be done in Python(should be present in the Ipython notebook), though they may be recreated in Tableau for better aesthetics in the PPT file.

Application Data: Cleaning

1. 22 columns were reduced to 73 as 49 columns had null values which were more than 32%.
2. Dropped 26 columns further as they were client documents and contact information. Hence the number of active columns became 47.
3. OCCUPATION_TYPE column had 31.3% null values marked them as "Unknown"
4. Replaced null values with median statistics for the following columns : EXT_SOURCE_3, EXT_SOURCE_2, AMT_GOODS_PRICE, AMT_ANNUITY
5. Replaced null values with mode statistics for the following columns : AMT_REQ_CREDIT_BUREAU_YEAR, AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_HOUR, NAME_TYPE_SUITE, OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE, CNT_FAM_MEMBERS, DAYS_LAST_PHONE_CHANGE

Application Data: Anomalies

1. Changed negative values to absolute for the following columns : DAYS_LAST_PHONE_CHANGE, DAYS_BIRTH, DAYS_REGISTRATION, DAYS_EMPLOYED, DAYS_ID_PUBLISH
2. CODE_GENDER column had XNA replaced them with mode statistics “F”
3. ORGANIZATION_TYPE column had XNA replaced them with “Unknow”
4. Following columns had outliers so did binning : AMT_INCOME_TOTAL, AMT_CREDIT, AMT_GOODS_PRICE, AMT_ANNUITY
5. Since DAYS_BIRTH has absolute values to predict the Age divided it by 365 . Then did binning to this column too.
6. Added 6 extra columns: AGE, AGE_RANGE, AMT_INCOME_TOTAL_RANGE, AMT_CREDIT_RANGE, AMT_GOODS_PRICE_RANGE, AMT_ANNUITY_RANGE
7. Final total number of rows and columns is : 307511 rows and 53 columns

Application Data: Analysis

TARGET is the dependent variable as it acts as first column to be checked if its 0 or 1 . If 0 then client with payment difficulties and if 0 then client with no payment difficulties.

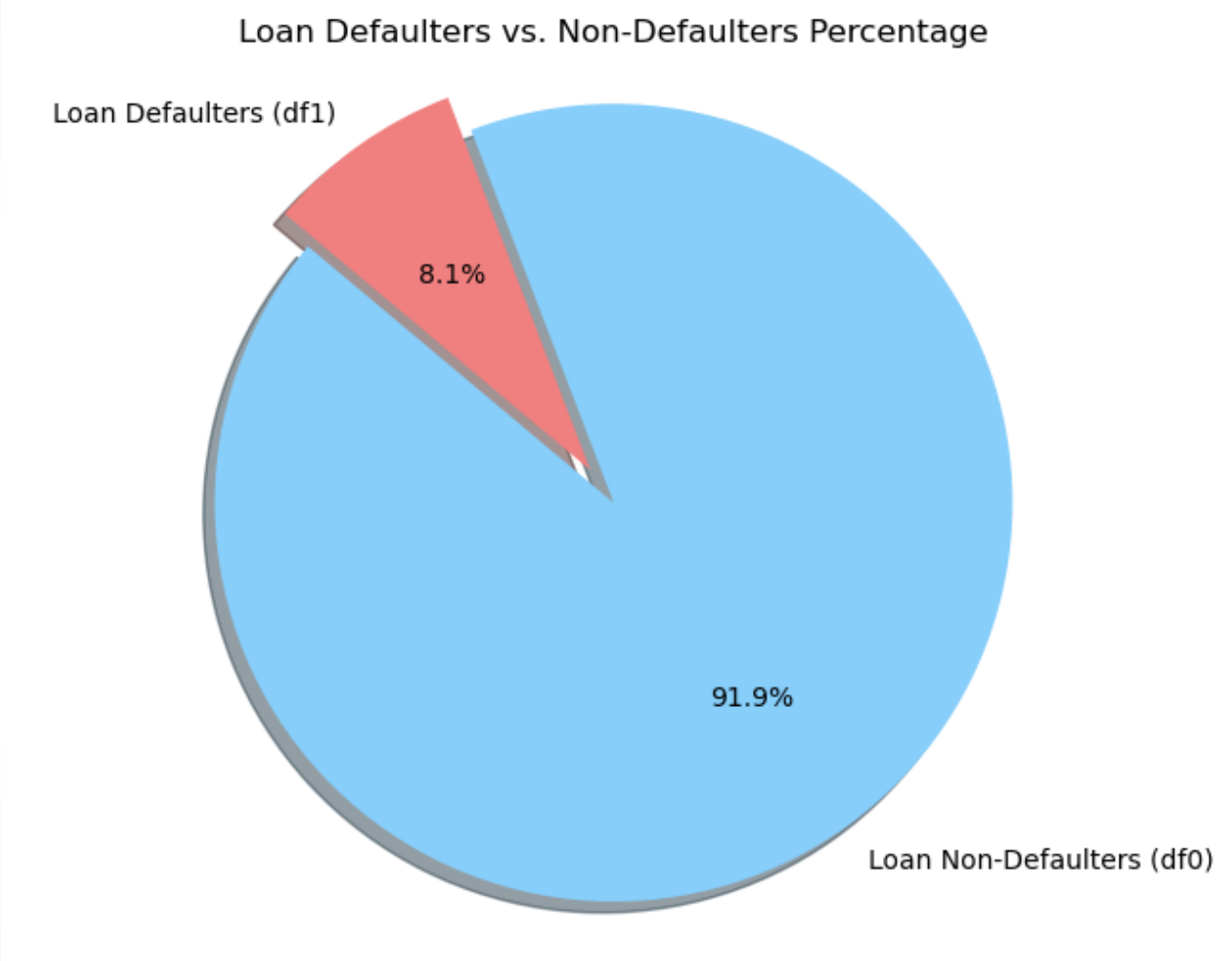
Diving the application data (df_app) into two data frames df0 (TARGET = 1 , client with payment difficulties)

df1(TARGET = 0 , all others)

Total rows with TARGET = 0 is 282,686 (Non-Defaulters)

Total rows with TARGET = 1 is 24,825(Defaulters)

Hence, we see from the graph that there are **8.1% population who are Defaulters.**

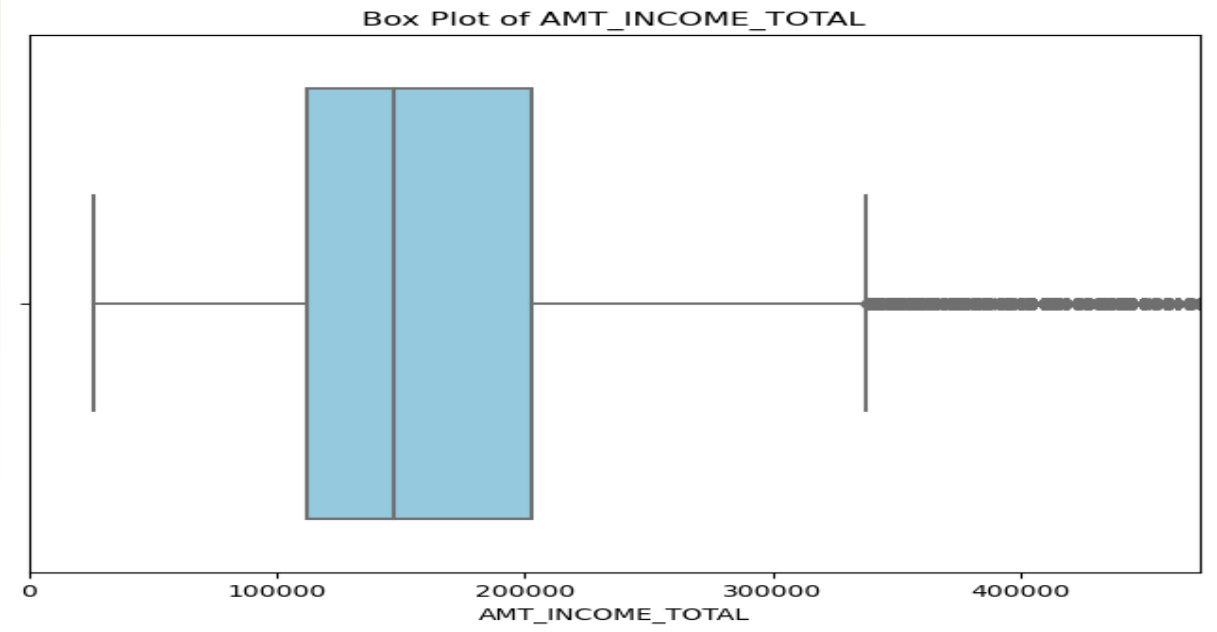


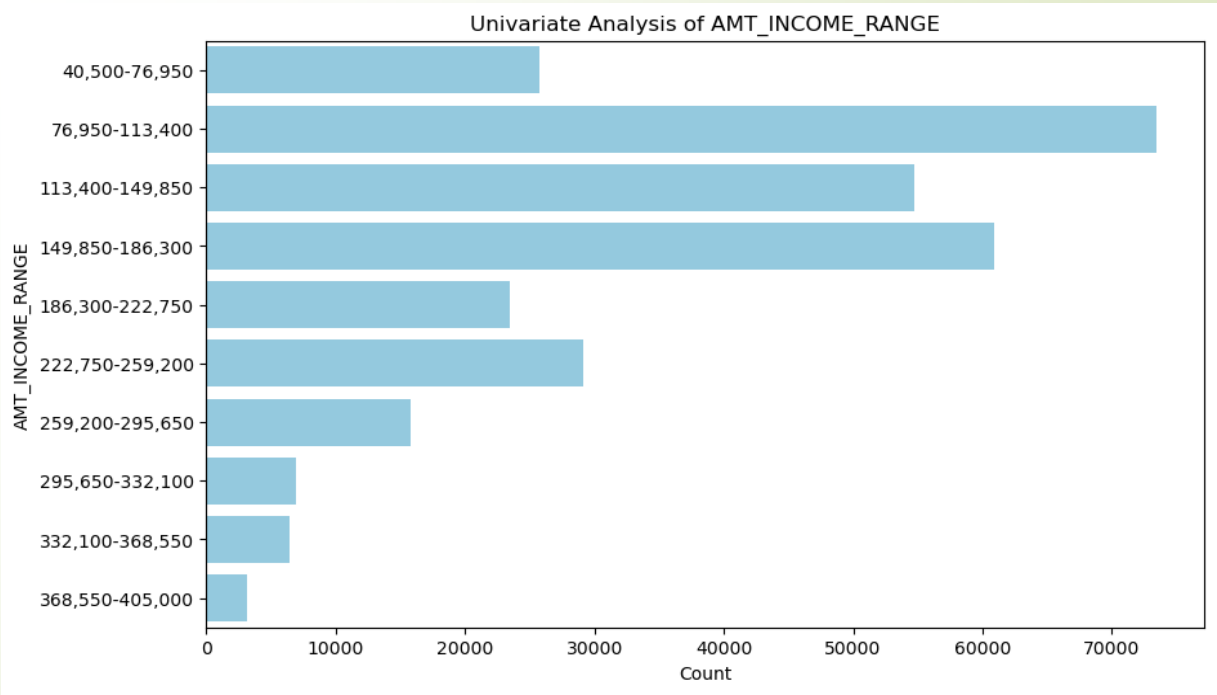
Univariate Analysis

AMT_INCOME:

As the raw data was skewed towards 75th percentile and has outliers , binning gave it more meaningful information .

1. We see a greater number of clients were in the income range 76 to 113K
2. Second largest were in the income range 149 to 186K followed by 113 to 149K



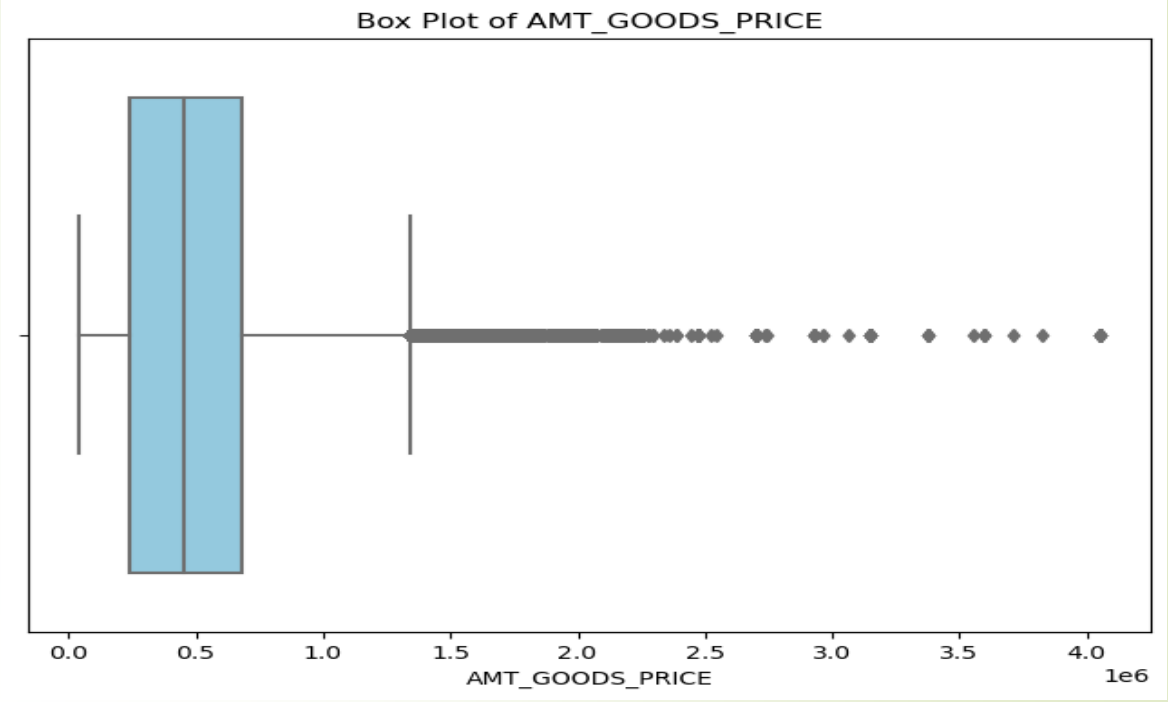


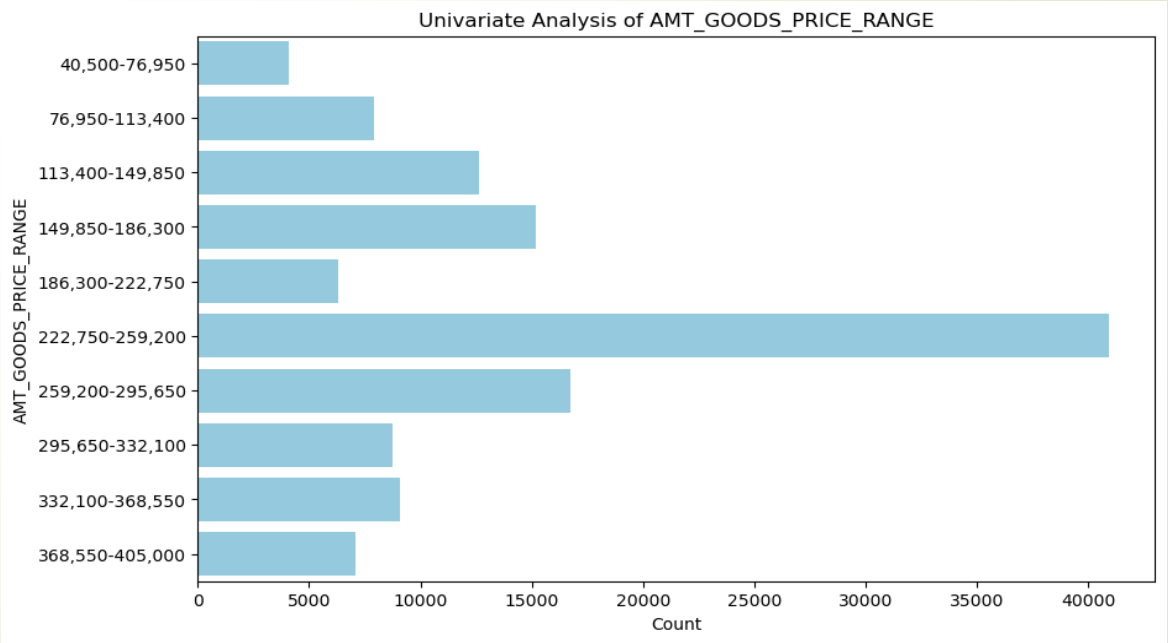
Univariate Analysis

GOODS_PRICE:

As the raw data was skewed towards 75th percentile and has outliers, binning gave it more meaningful information .

1. We see a greater number of clients took loan for the goods which were priced in the range of 222 to 259K
2. Second largest were in the range of 259 to 295K and 149 to 186K
3. Third was in the range of 113 to 149K



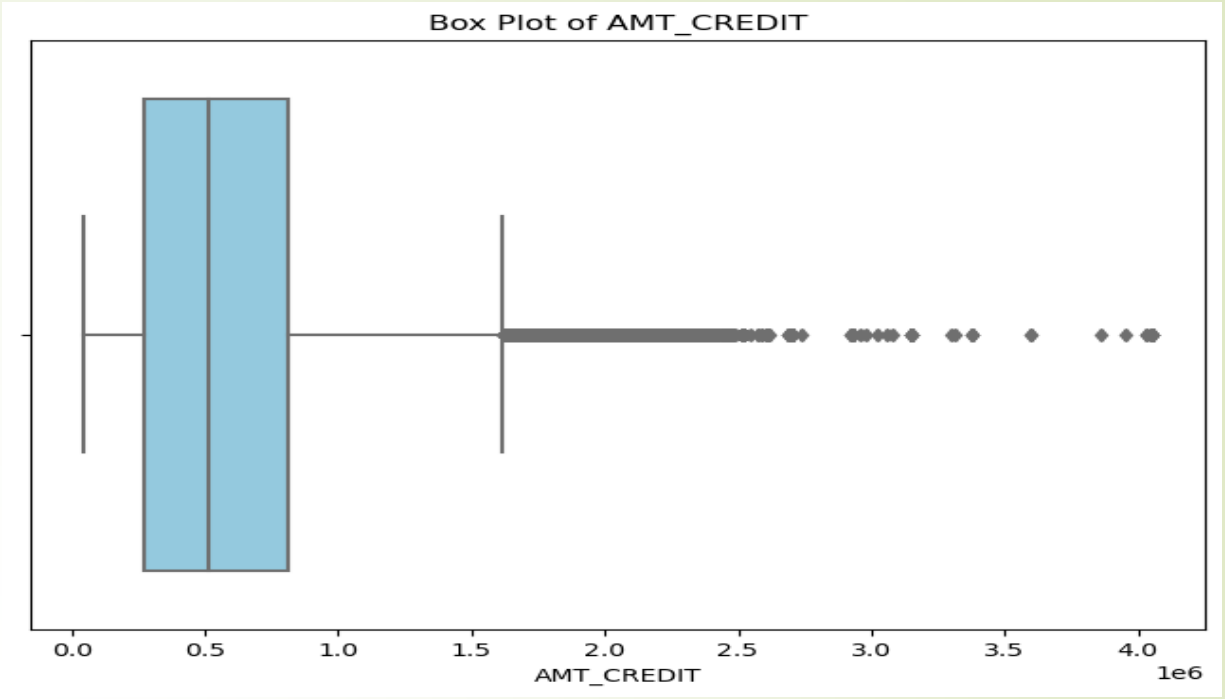


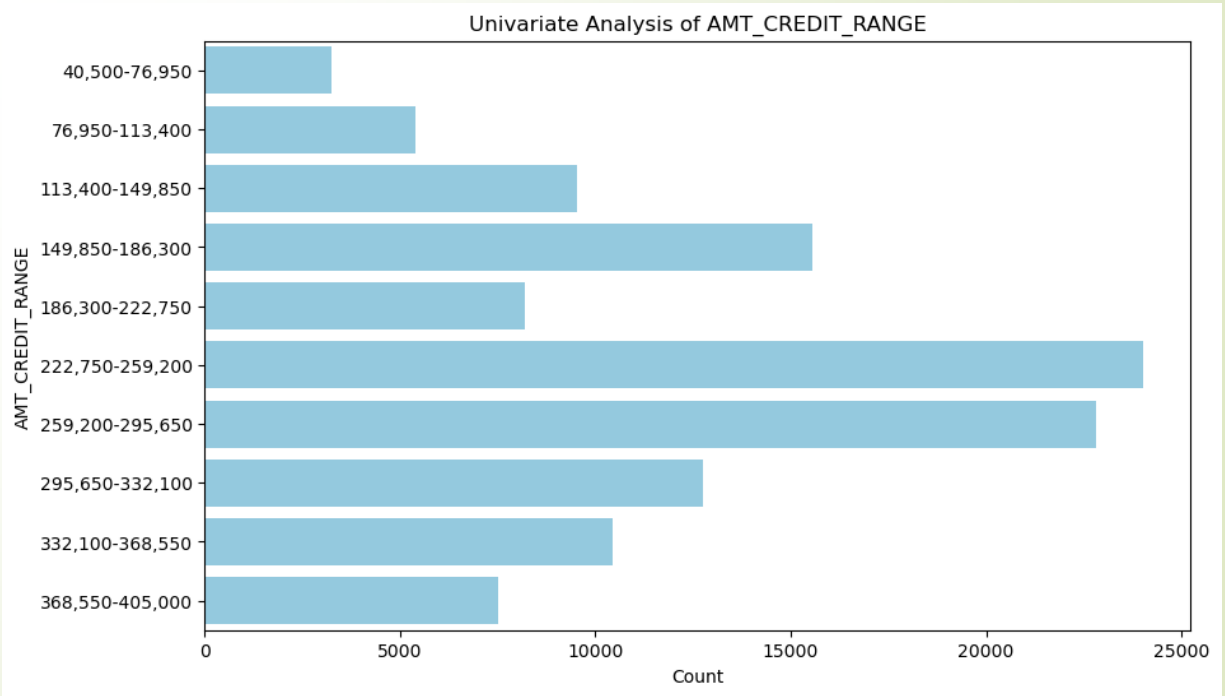
Univariate Analysis

AMT_CREDIT:

As the raw data was skewed towards 75th percentile and has outliers , binning gave it more meaningful information .

1. We see a greater number of clients took credit in the range of 222 to 259K
2. Second largest were in the range 259 to 295K followed by 140 to 186K



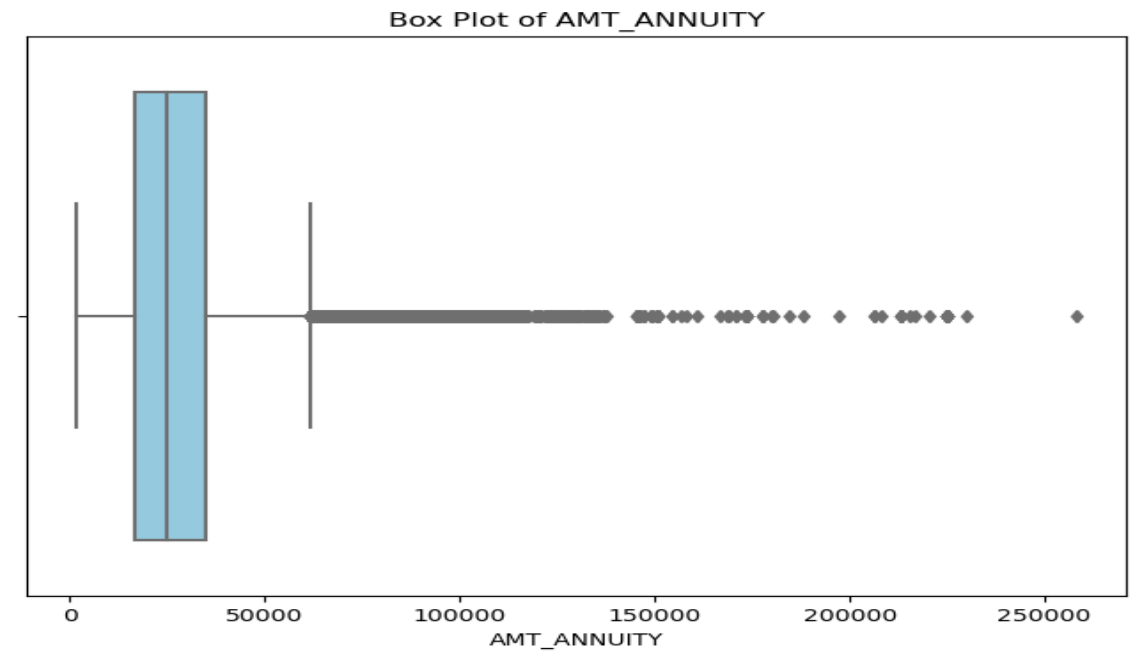


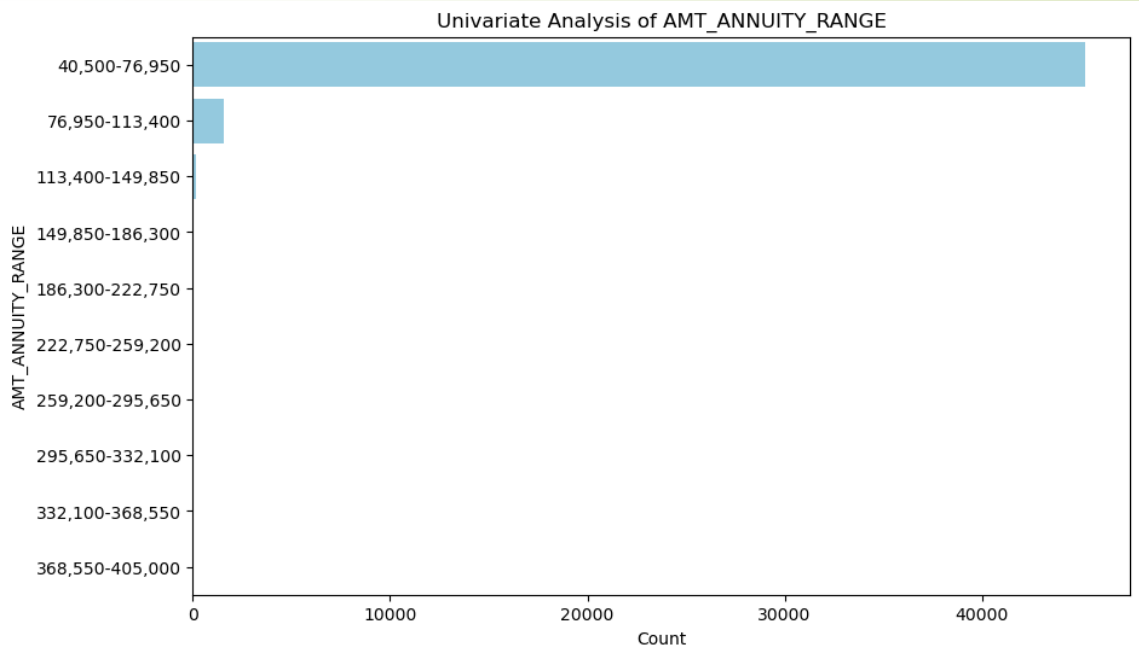
Univariate Analysis

AMT_ANNUITY:

As the raw data was skewed towards 75th percentile and has outliers, binning gave it more meaningful information .

1. We see a greater number of clients were in the loan annuity range 40 to 76K followed by 76 to 113K



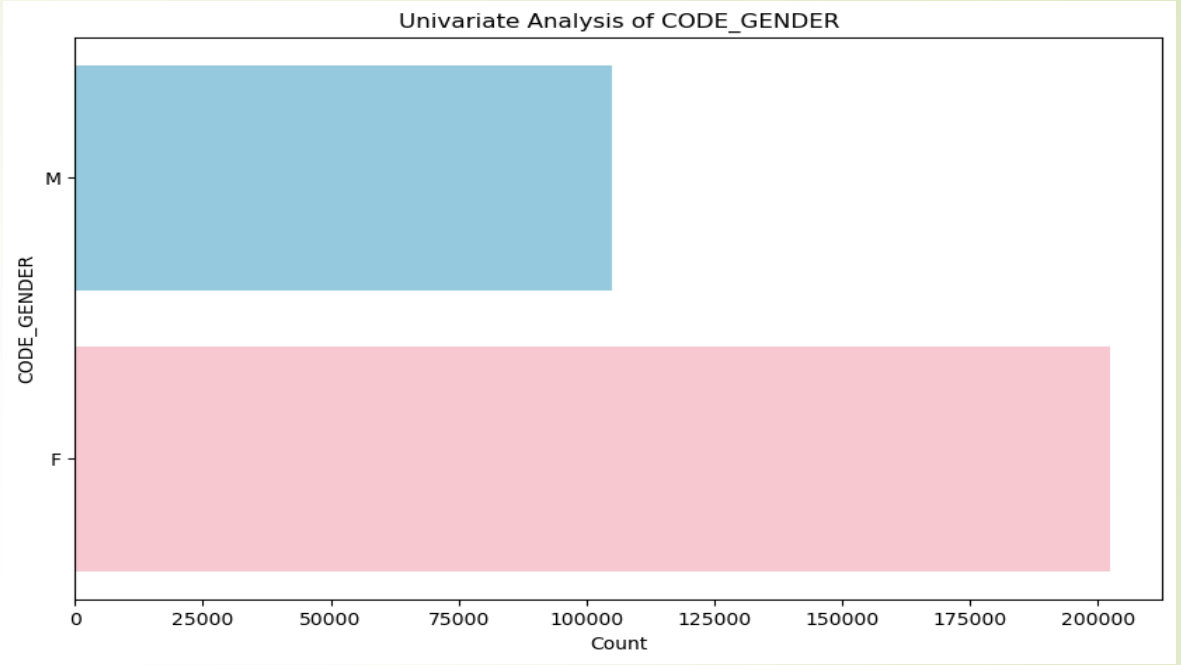


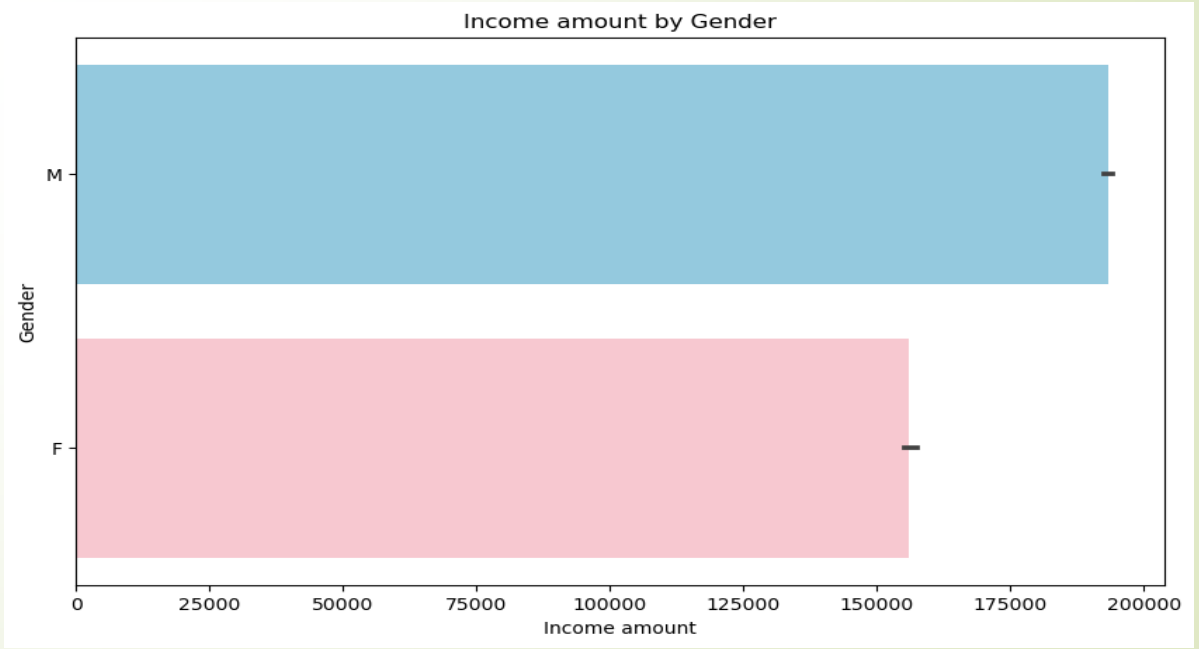
Univariate Analysis

CODE_GENDER:

We see more number of female population in the group then male who have taken loan.

But when compared with the income that both gender take , male population earns significantly higher than females.

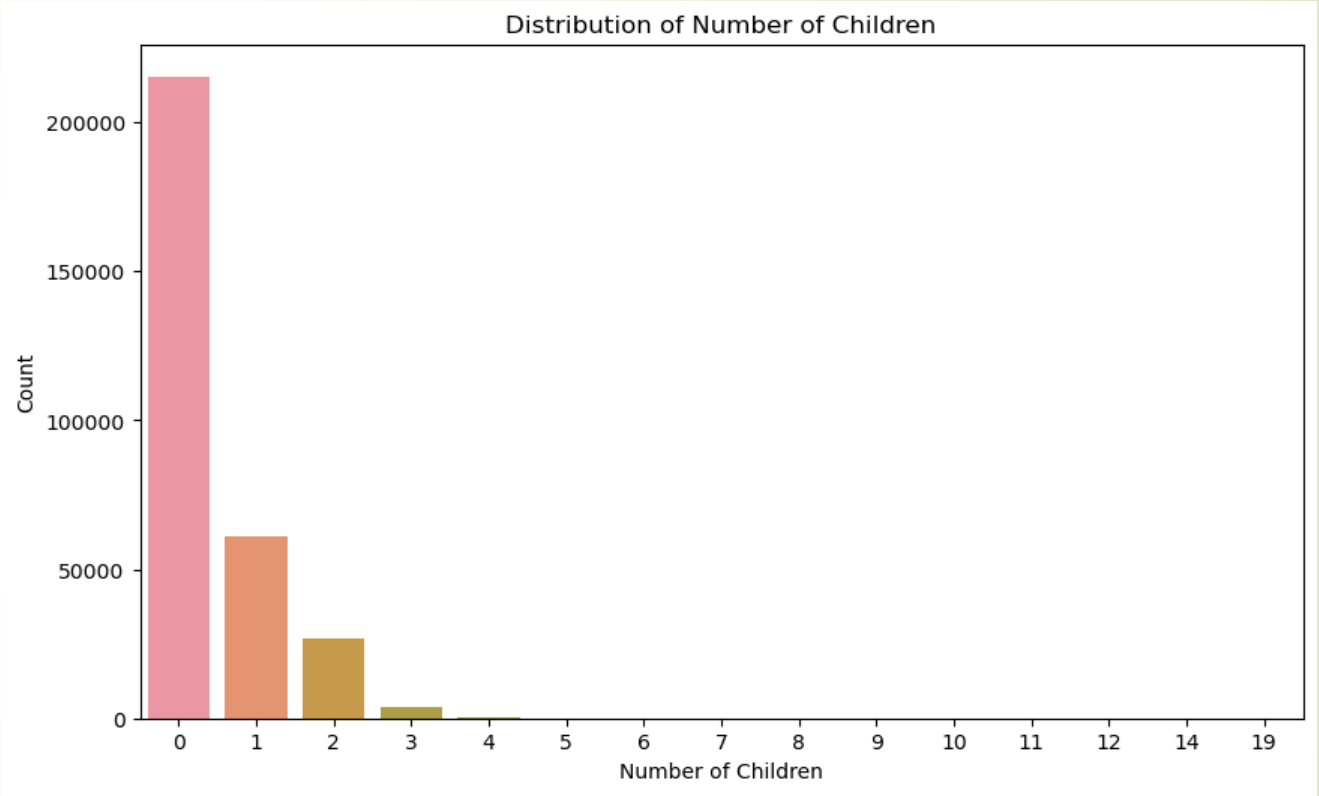




Univariate Analysis

CNT_CHILDREN:

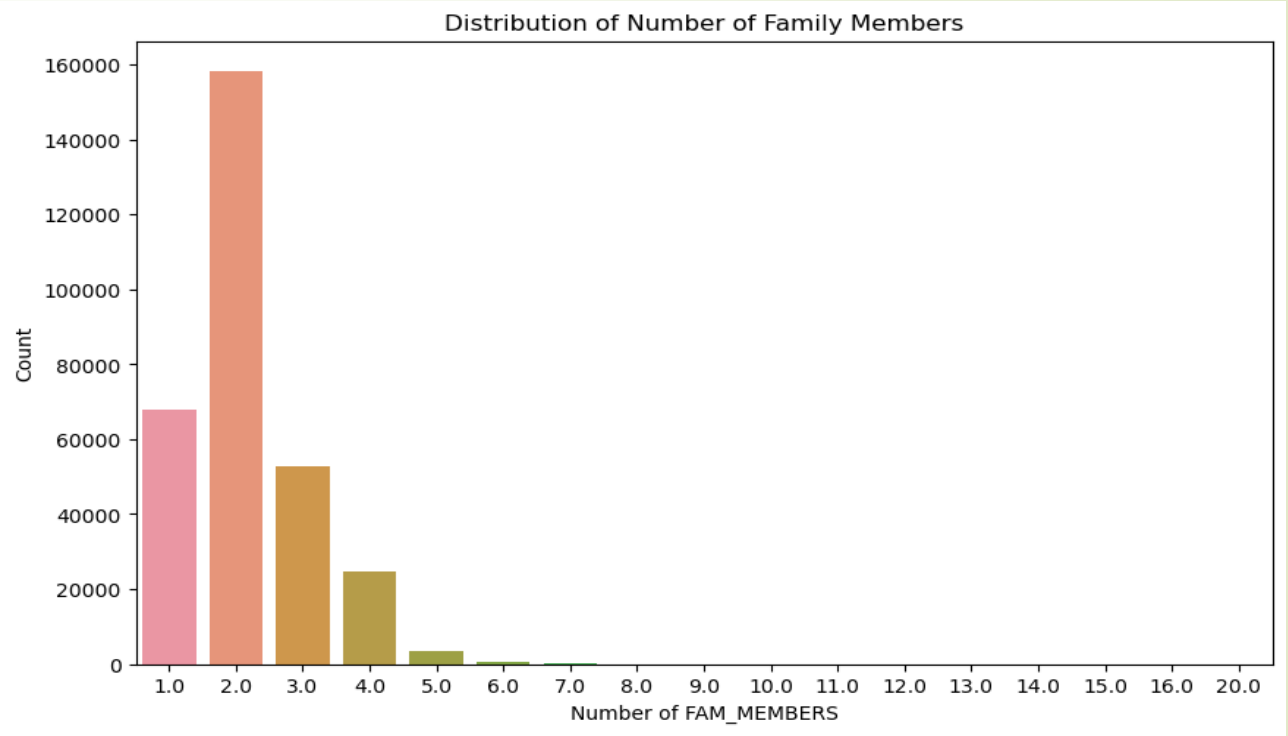
We see significantly high number of clients who do not have children.



Univariate Analysis

CNT_FAM_MEMBERS:

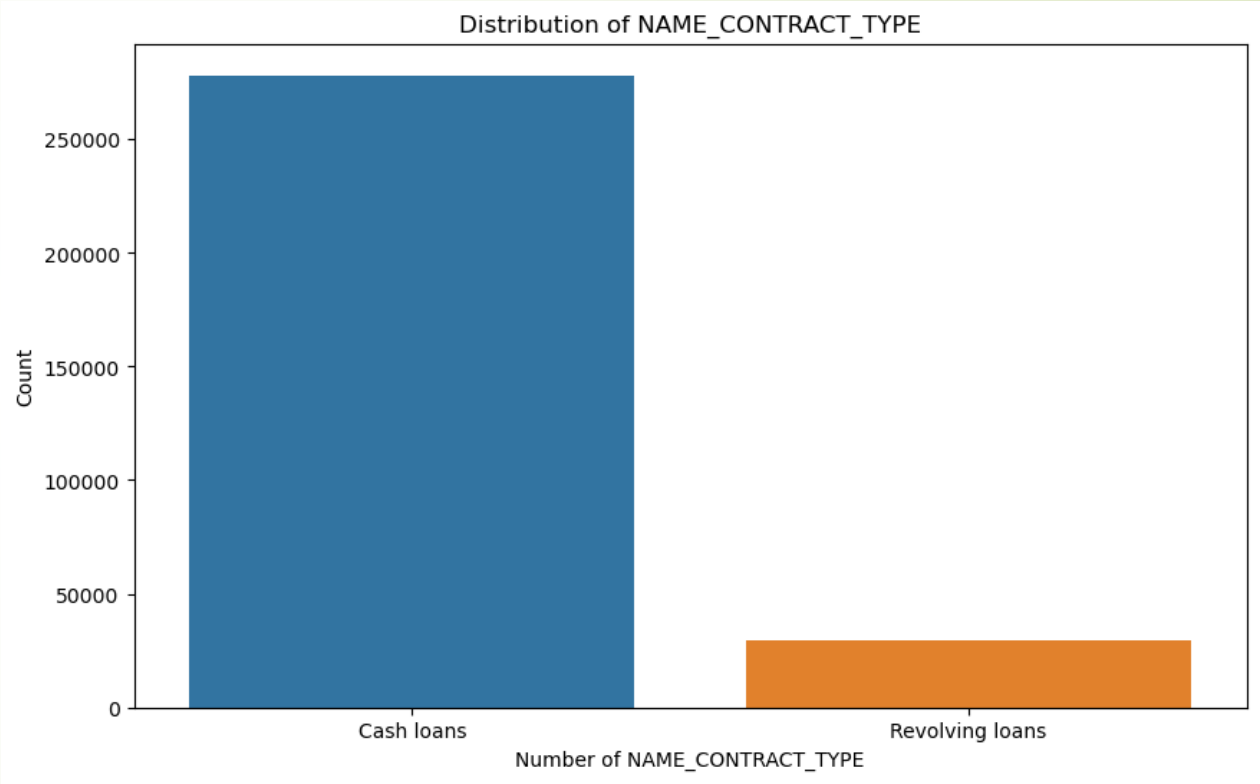
We see a significantly large number of clients who have 2 family members



Univariate Analysis

NAME_CONTRACT_TYPE:

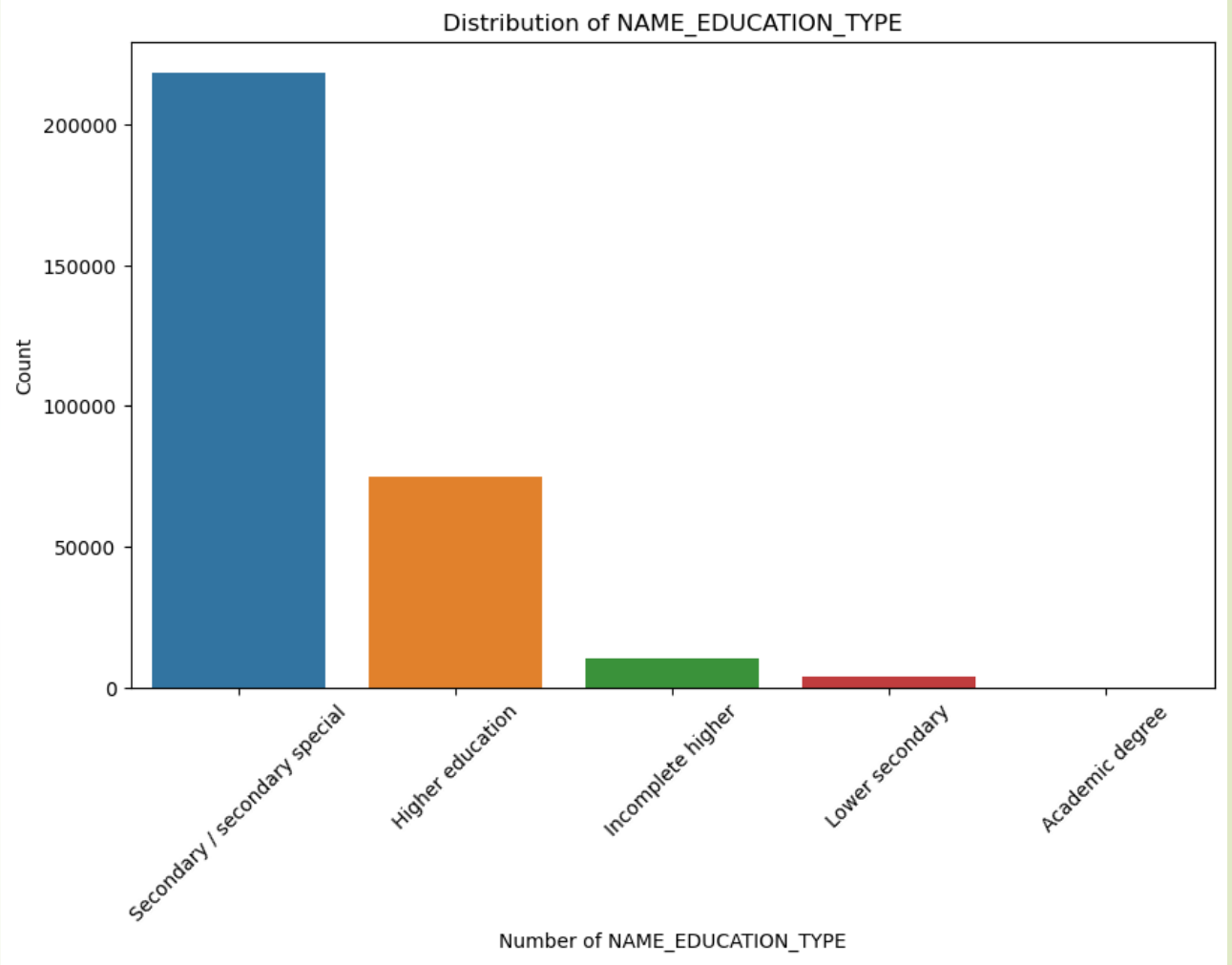
We see a significantly large number of clients who have availed Cash Loans



Univariate Analysis

NAME_EDUCATION_TYPE:

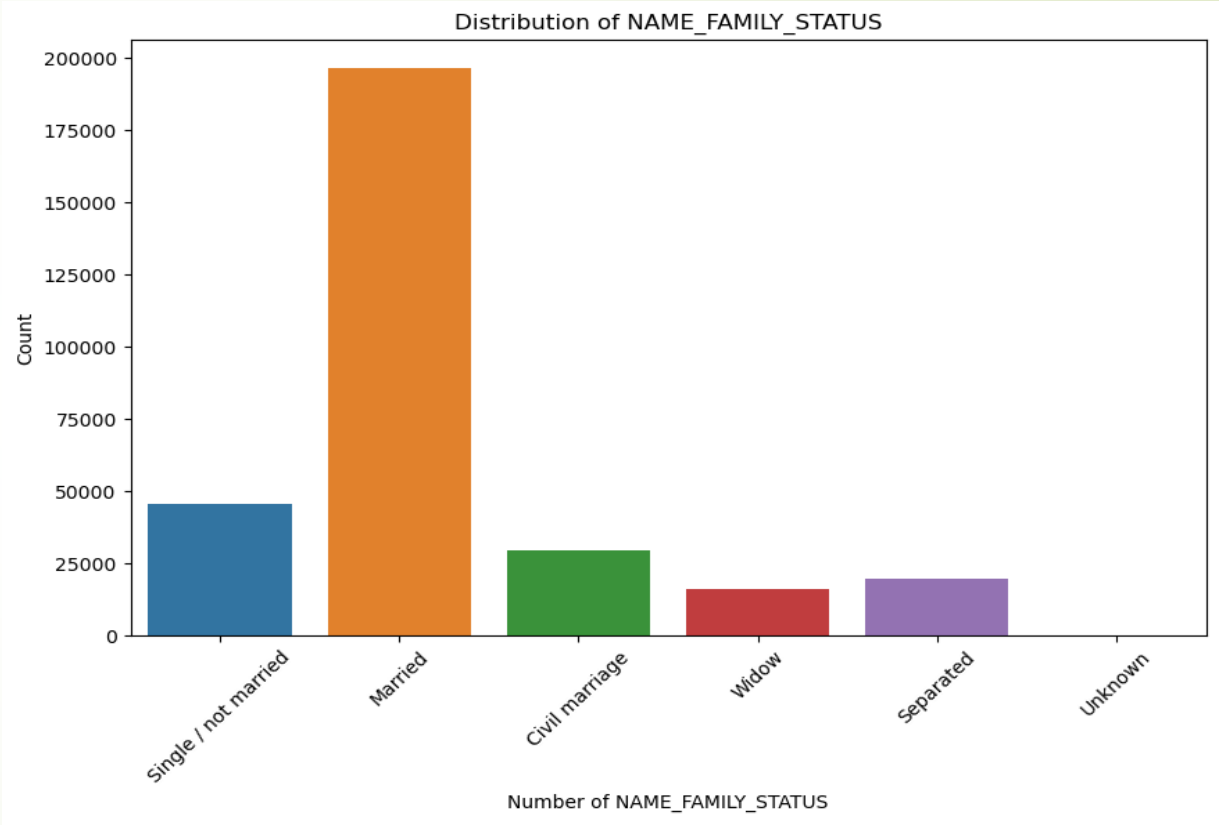
We see a significantly large number of clients who are Secondary/secondary special education group who have taken loan



Univariate Analysis

NAME_FAMILY_STATUS

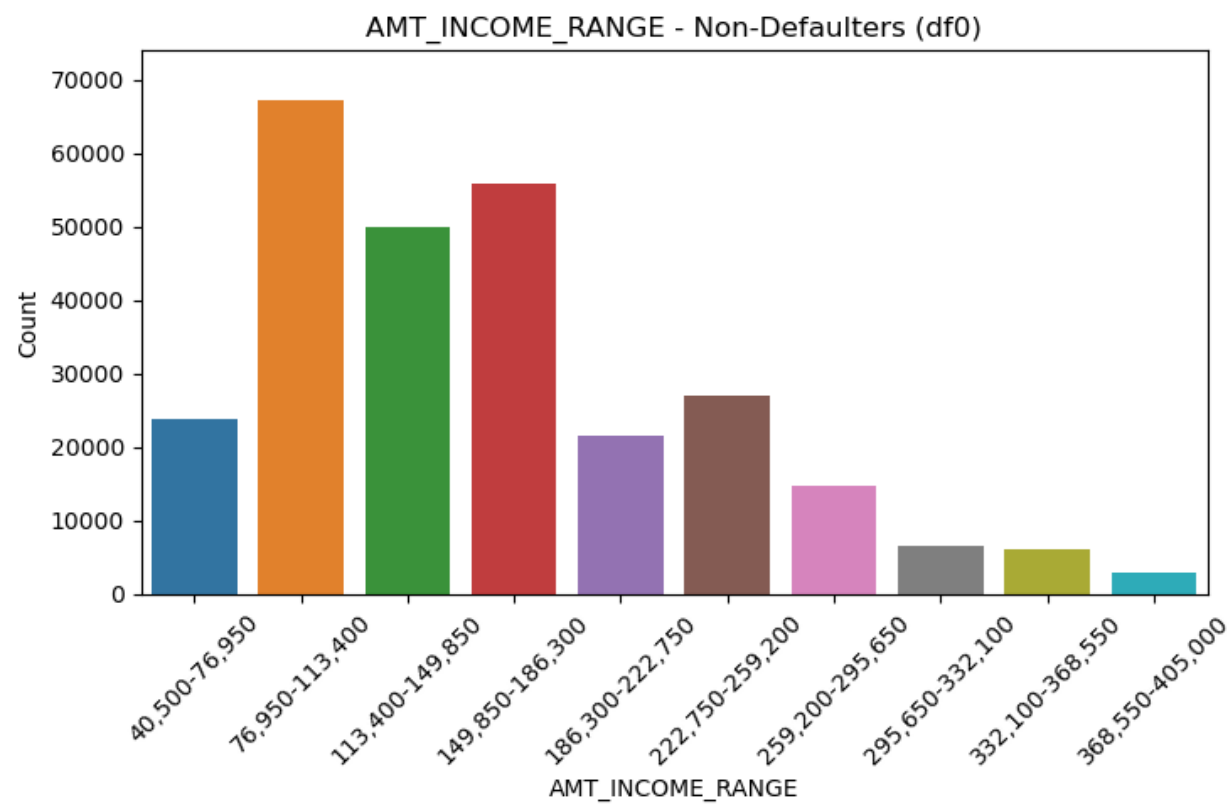
We see a significantly large number of clients who are from Married group have taken loan

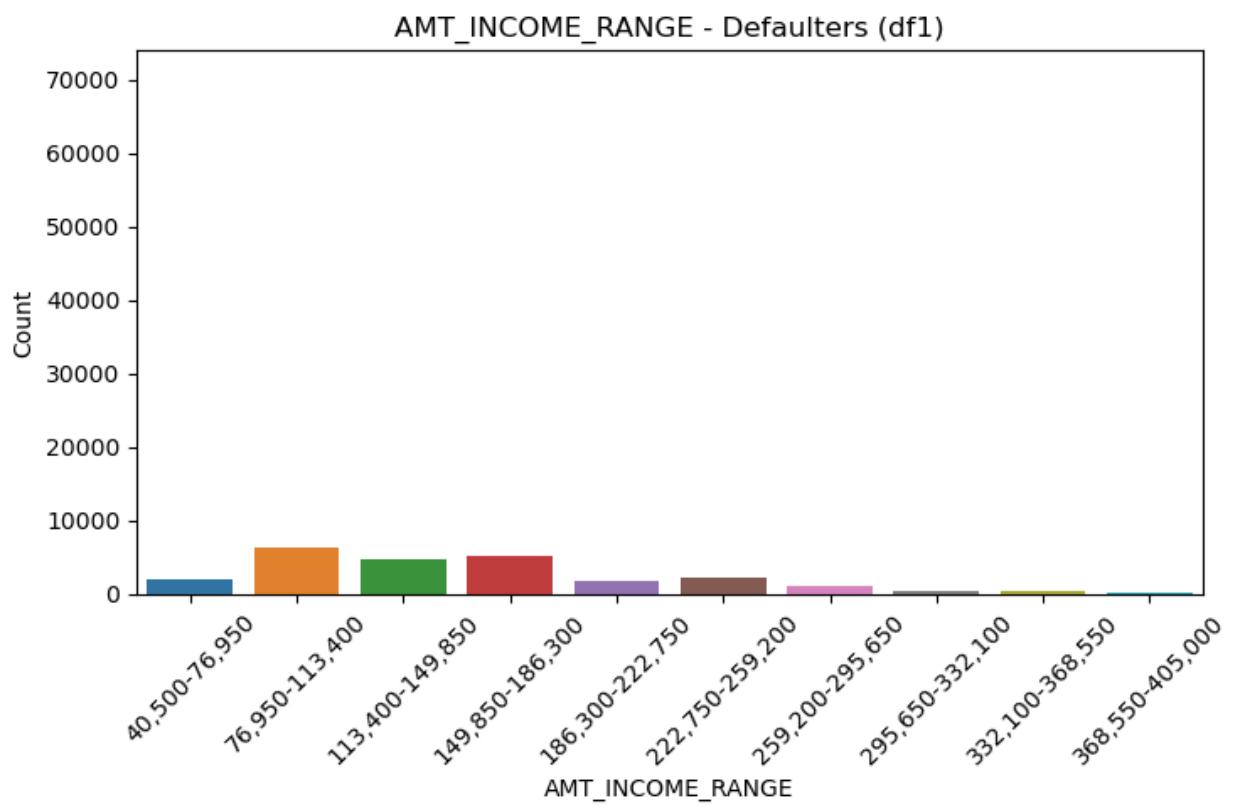


Bivariate Analysis

To cross check at columns from TARGET perspective(defaulters and non-defaulters)

- Both groups(defaulters & non-defaulters) have highest **INCOME** in the range of 76 to 113K

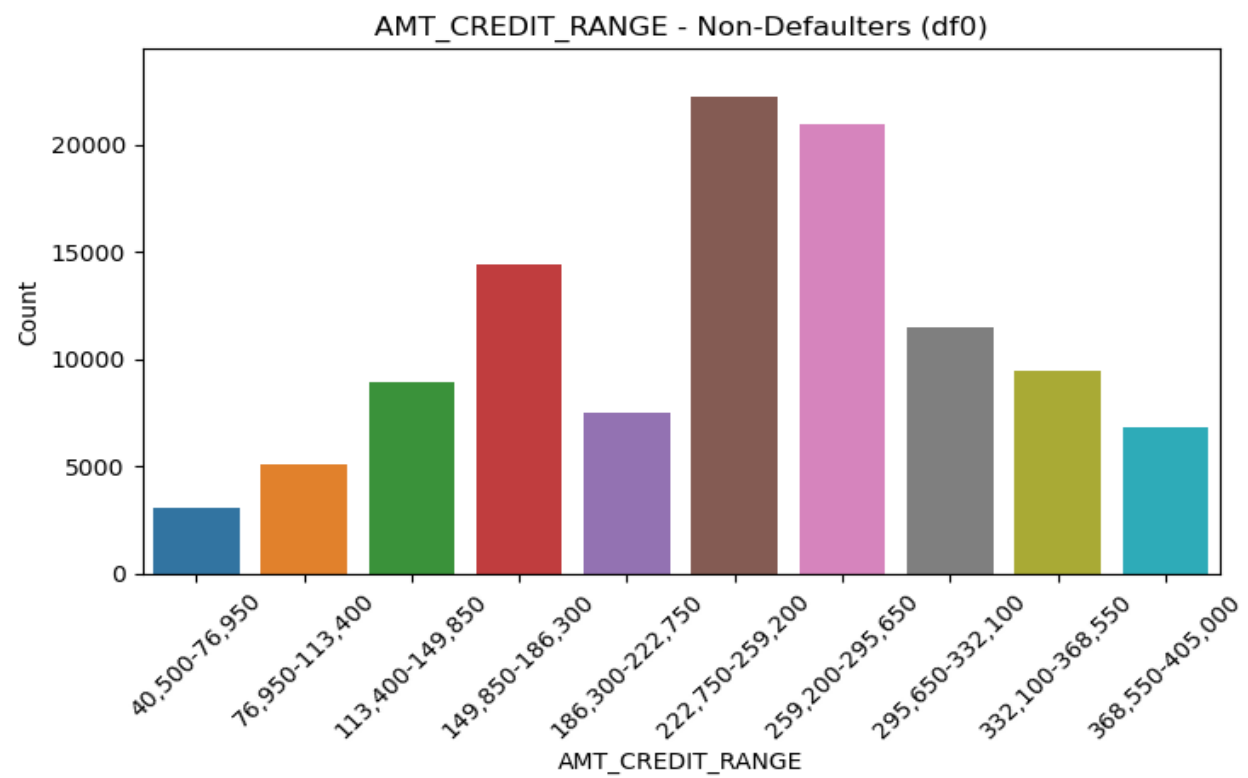


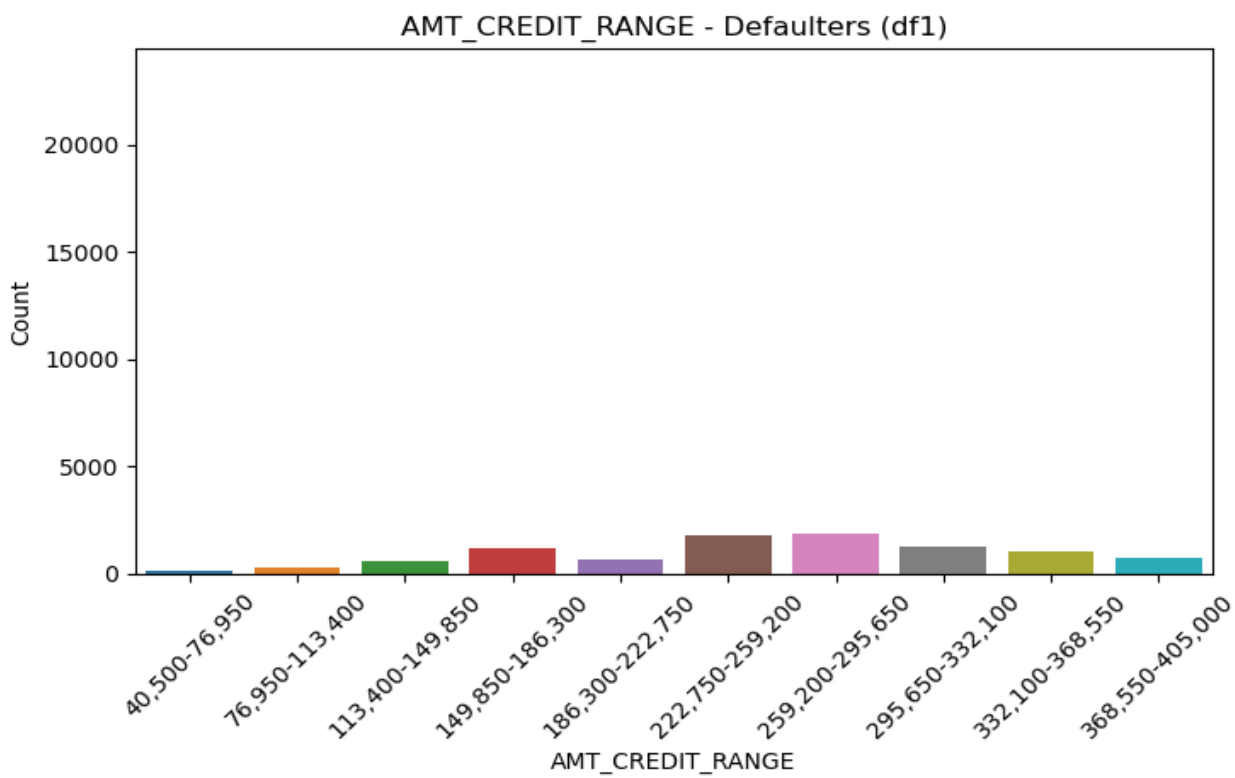


Bivariate Analysis

To cross check at columns from TARGET perspective(defaulters and non-defaulters)

- More number of non-defaulter clients have taken higher **CREDIT_AMT** then defaulters.

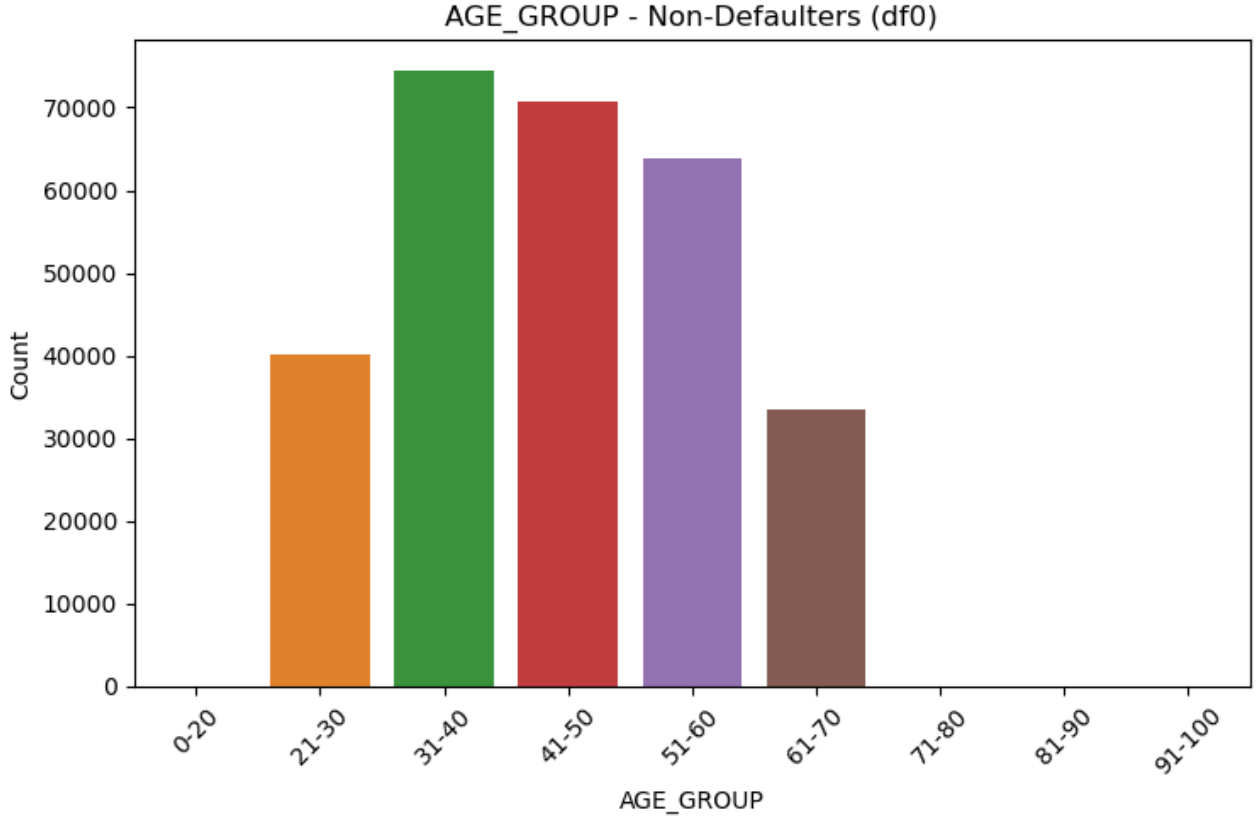


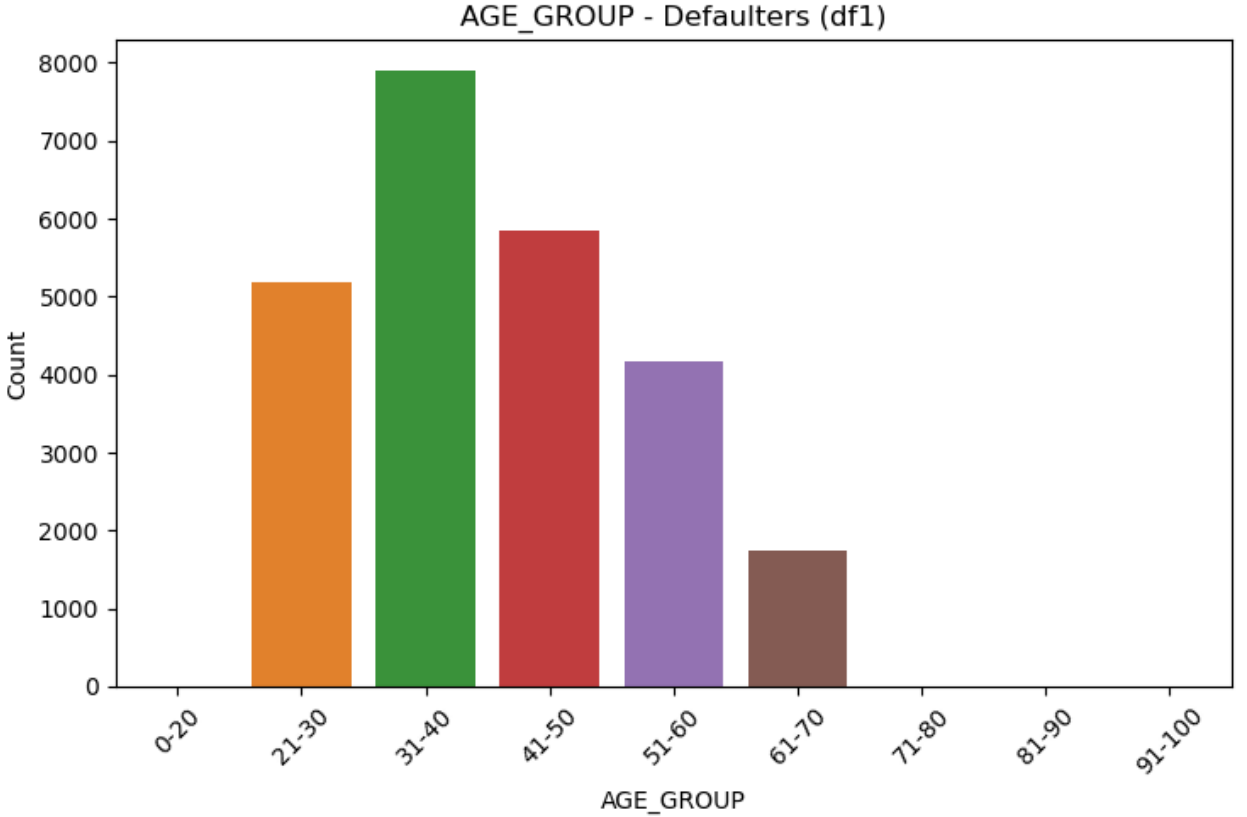


Bivariate Analysis

To cross check at columns from TARGET perspective(defaulters and non-defaulters)

- **31- to 40-year-olds** were higher defaulters then other age groups in "Defaulters(df1)

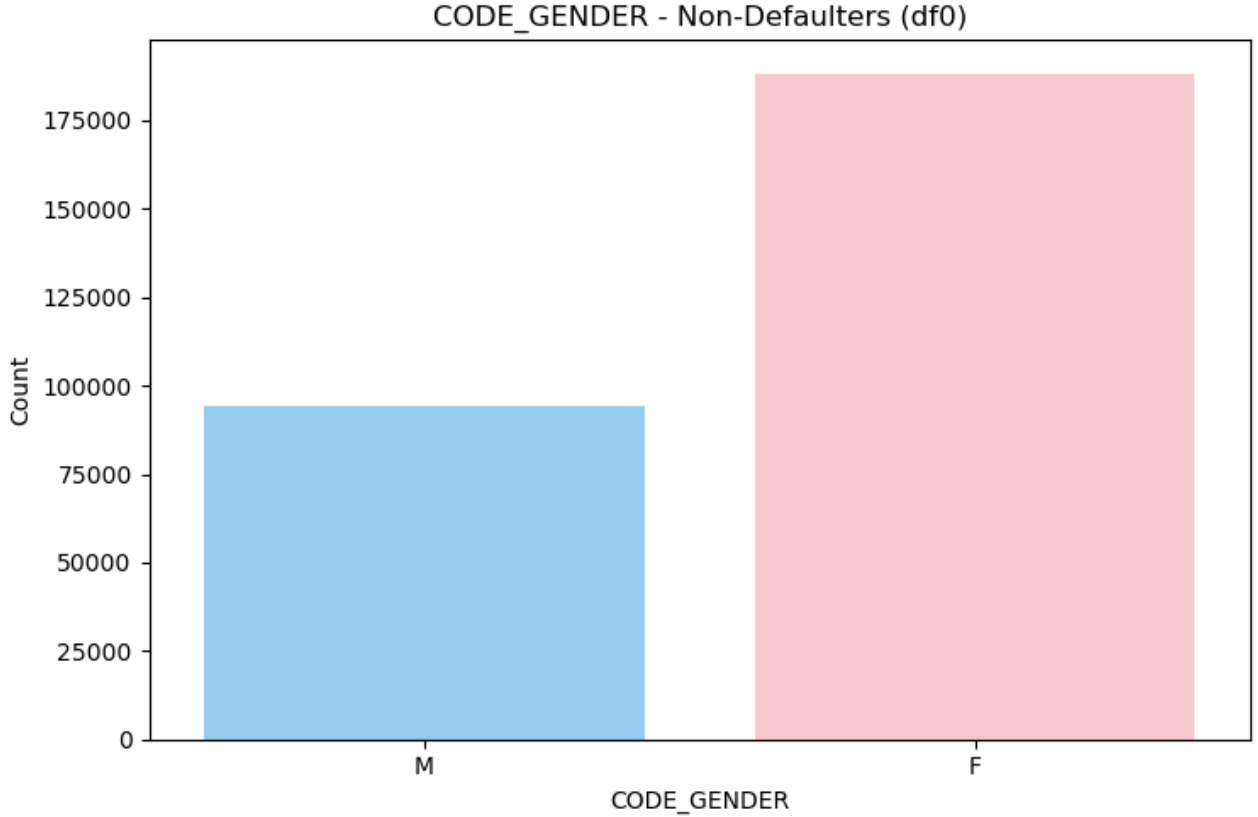


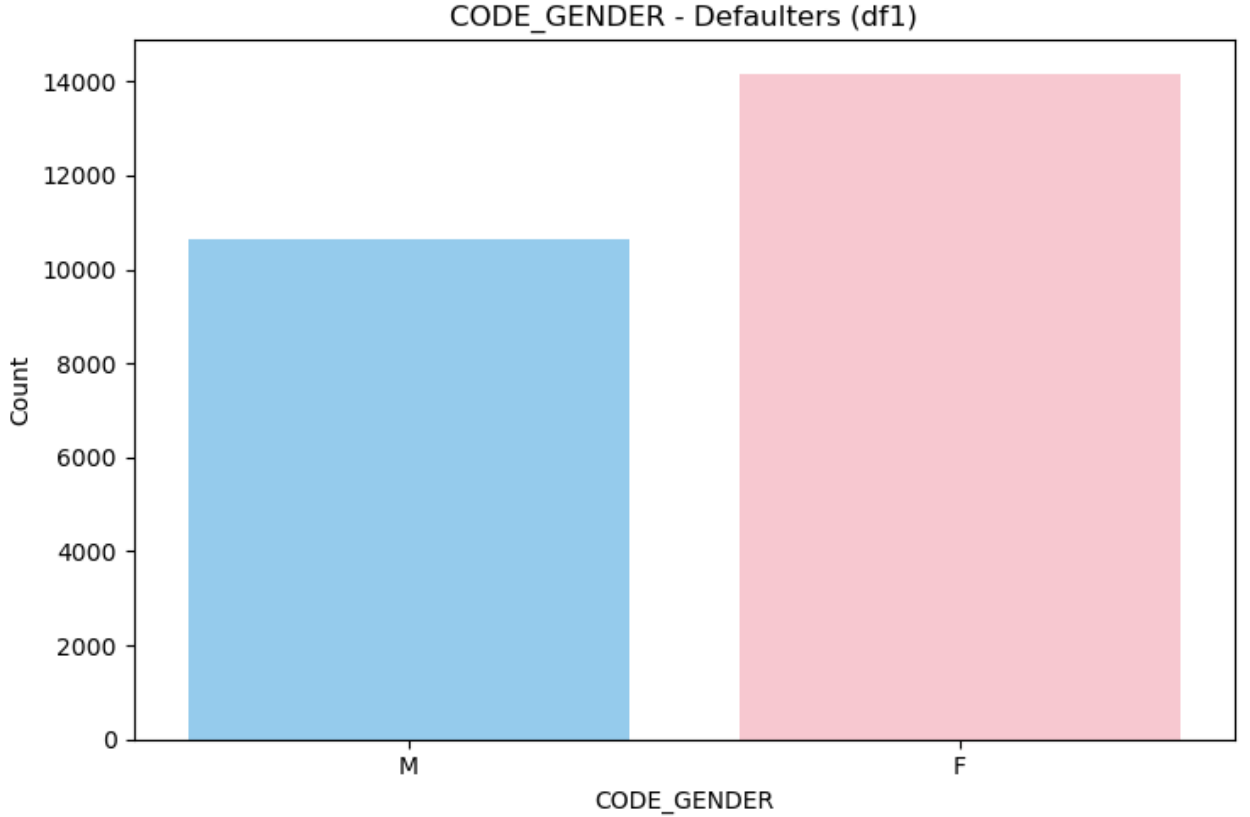


Bivariate Analysis

To cross check at columns from TARGET perspective(defaulters and non-defaulters)

- In both the TARGET groups, **Female** population was highest.

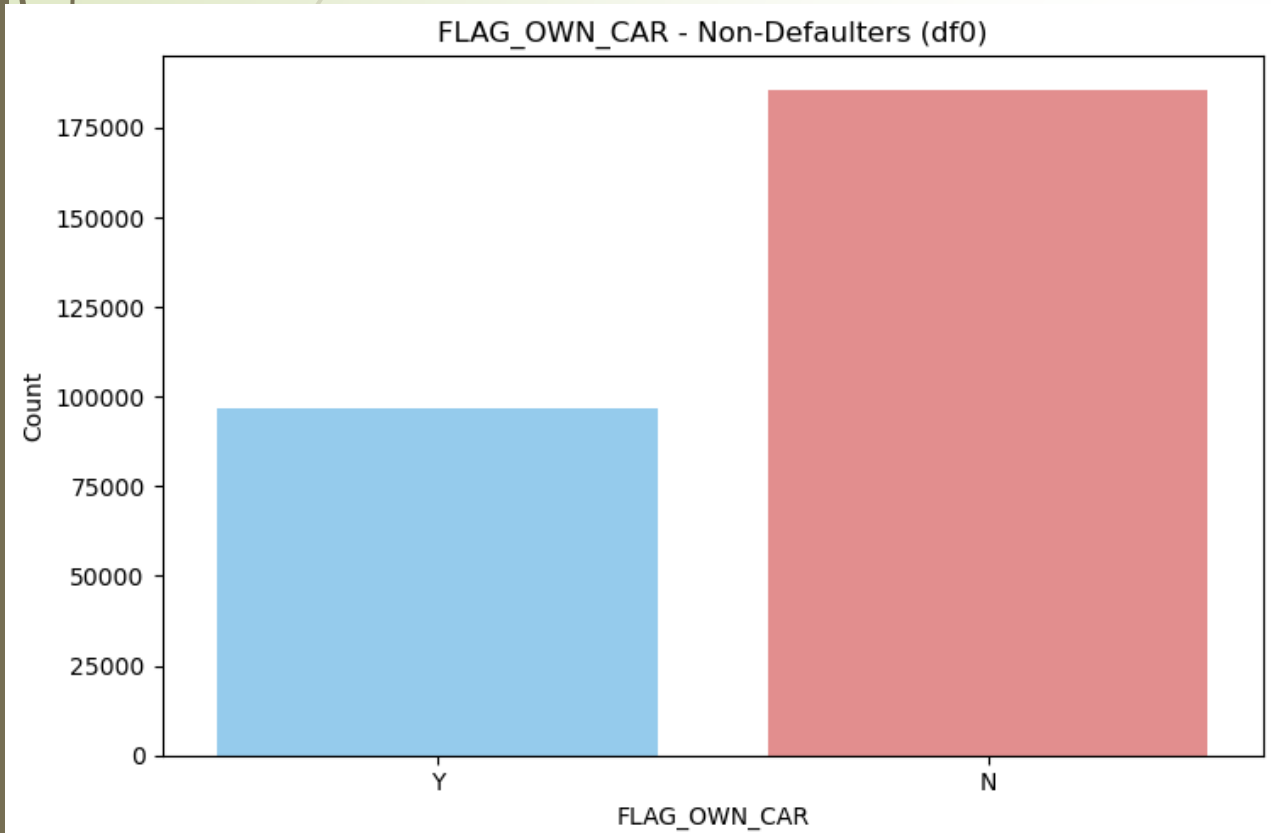


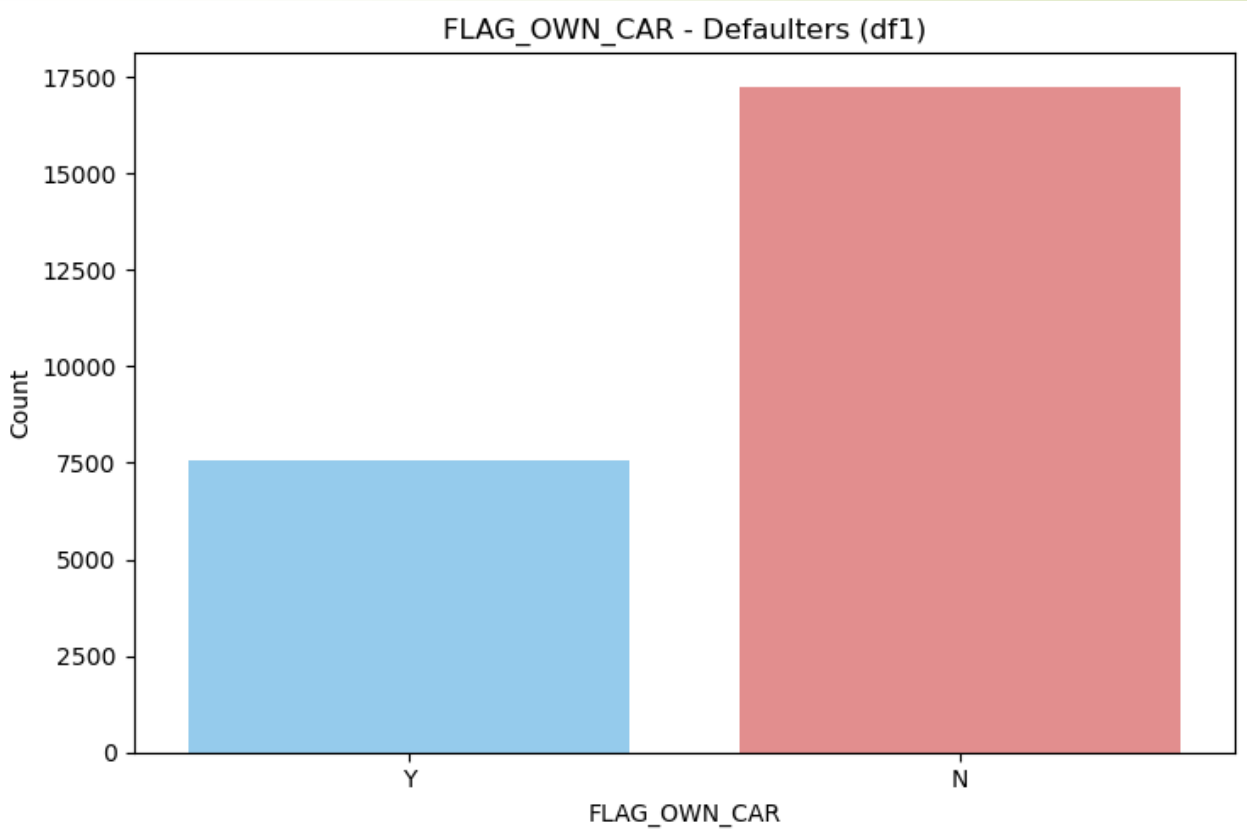


Bivariate Analysis

To cross check at columns from TARGET perspective(defaulters and non-defaulters)

- In both the TARGET groups, most of them did not own a **car** while applying for the loan.

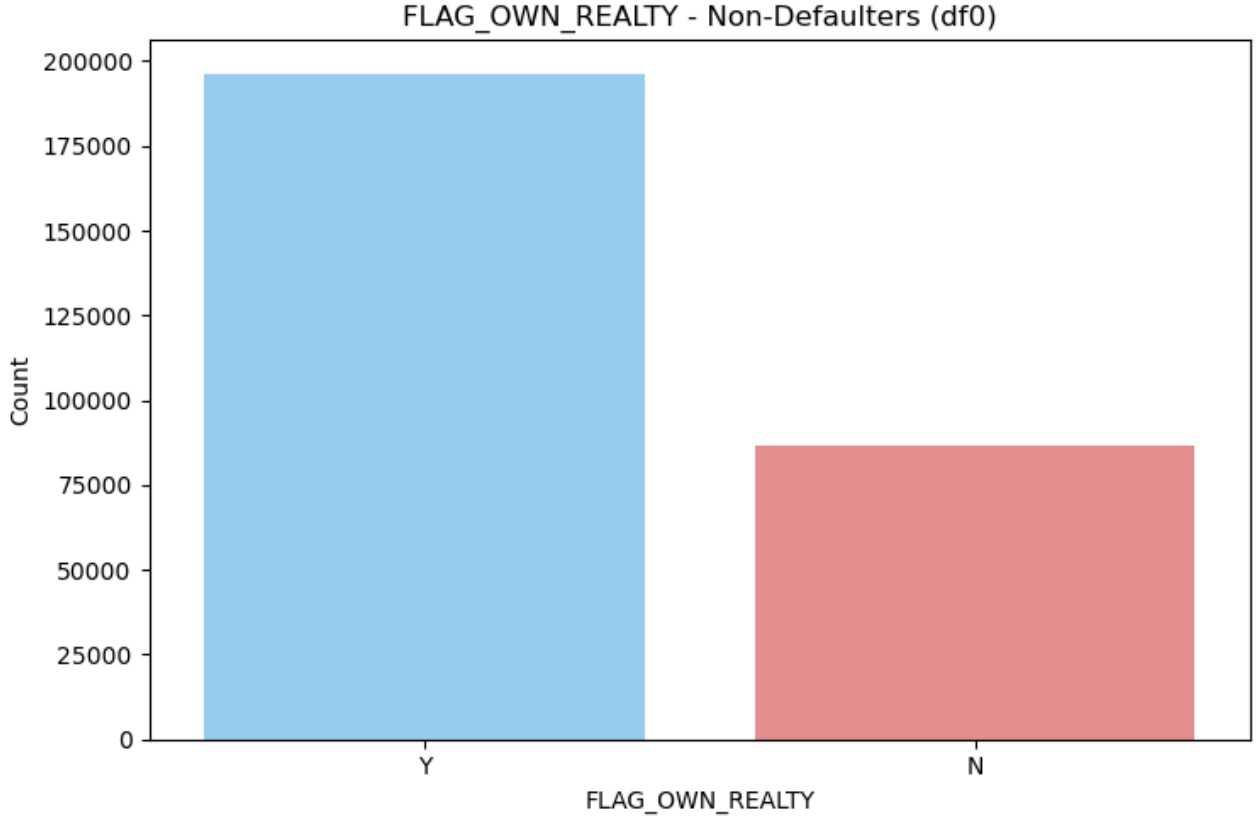


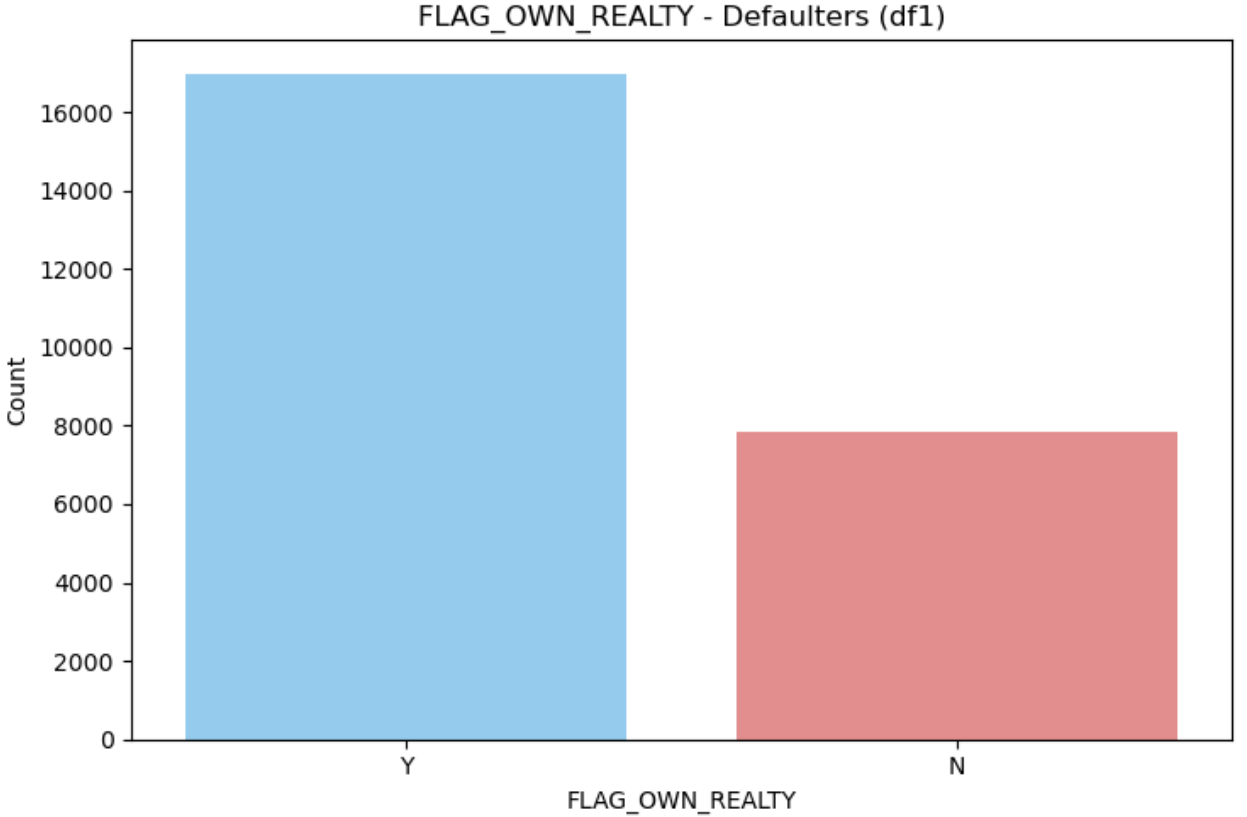


Bivariate Analysis

To cross check at columns from TARGET perspective(defaulters and non-defaulters)

- In both the TARGET groups, most of them owned a **house or flat** while applying for the loan.

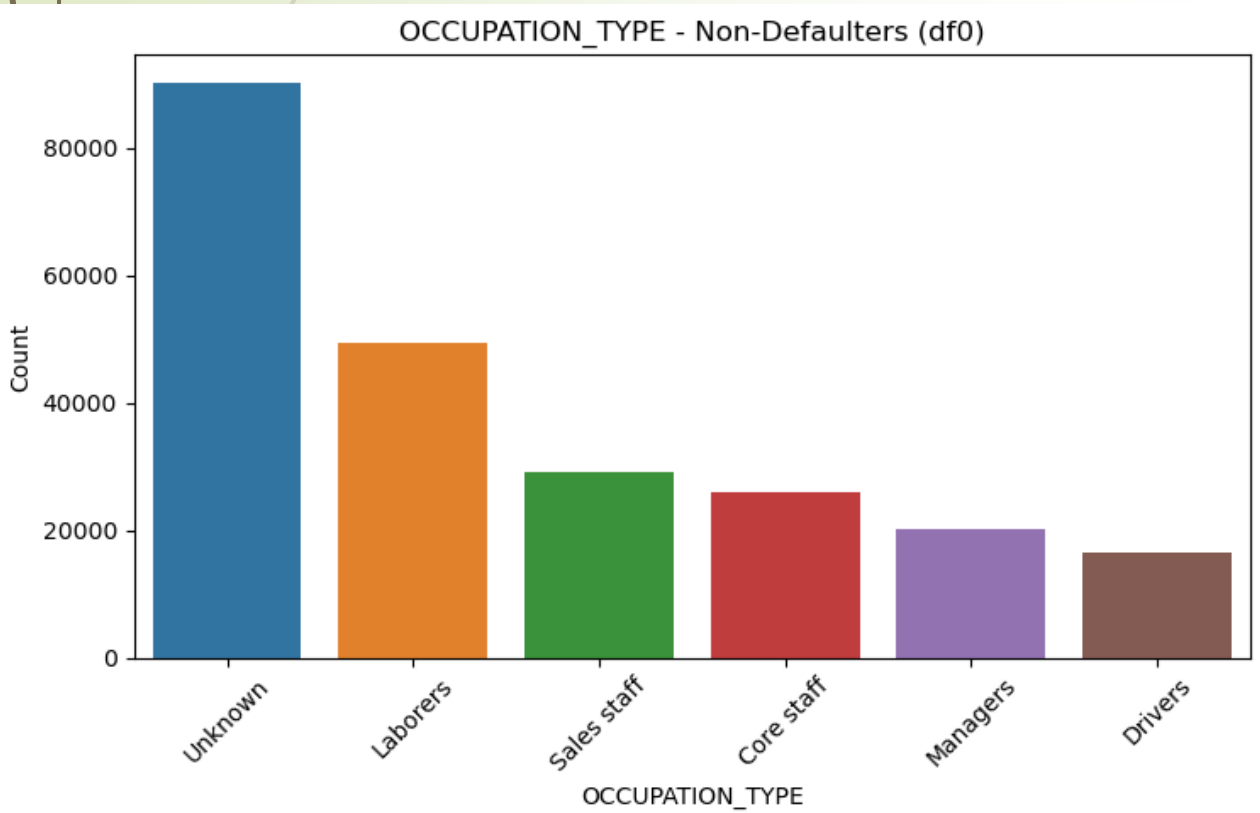


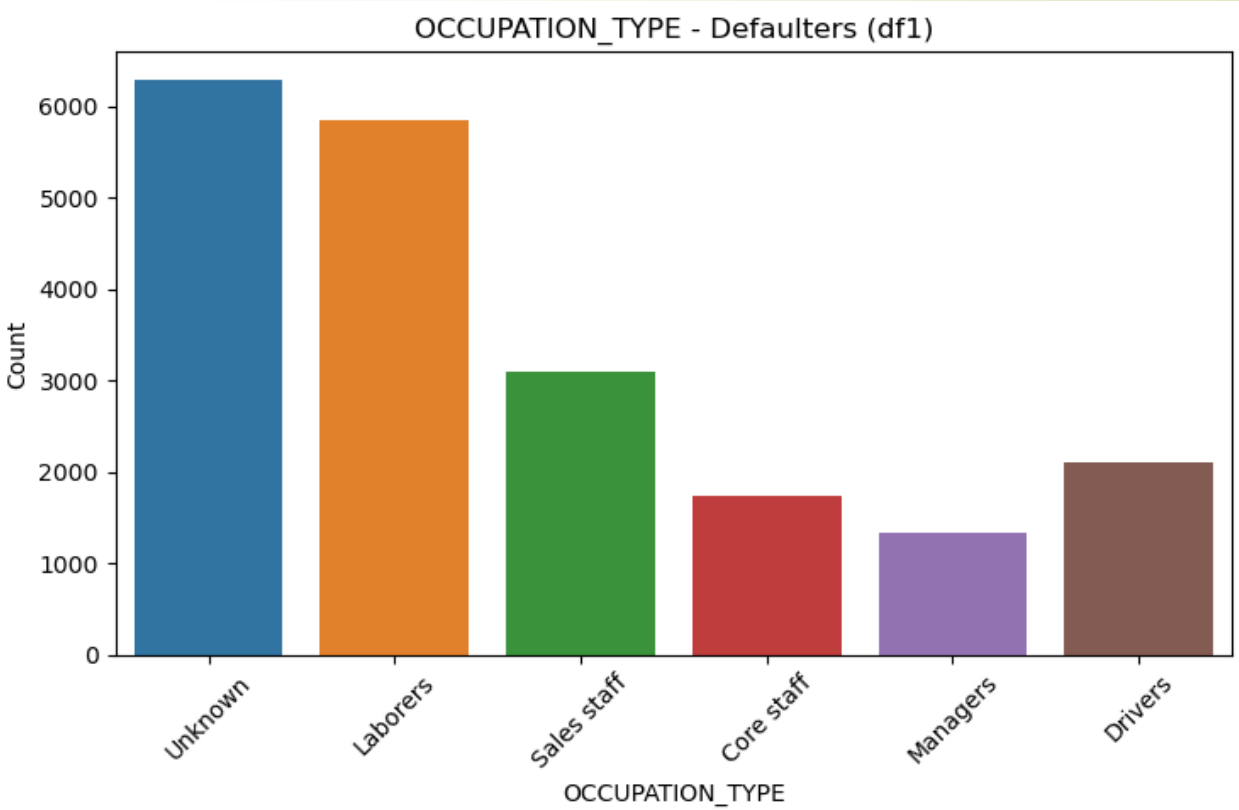


Bivariate Analysis

To cross check at columns from TARGET perspective(defaulters and non-defaulters)

- In both the TARGET groups , "**OCCUPATION_TYPE**" column saw **Laborers** who were the one's who took more loan

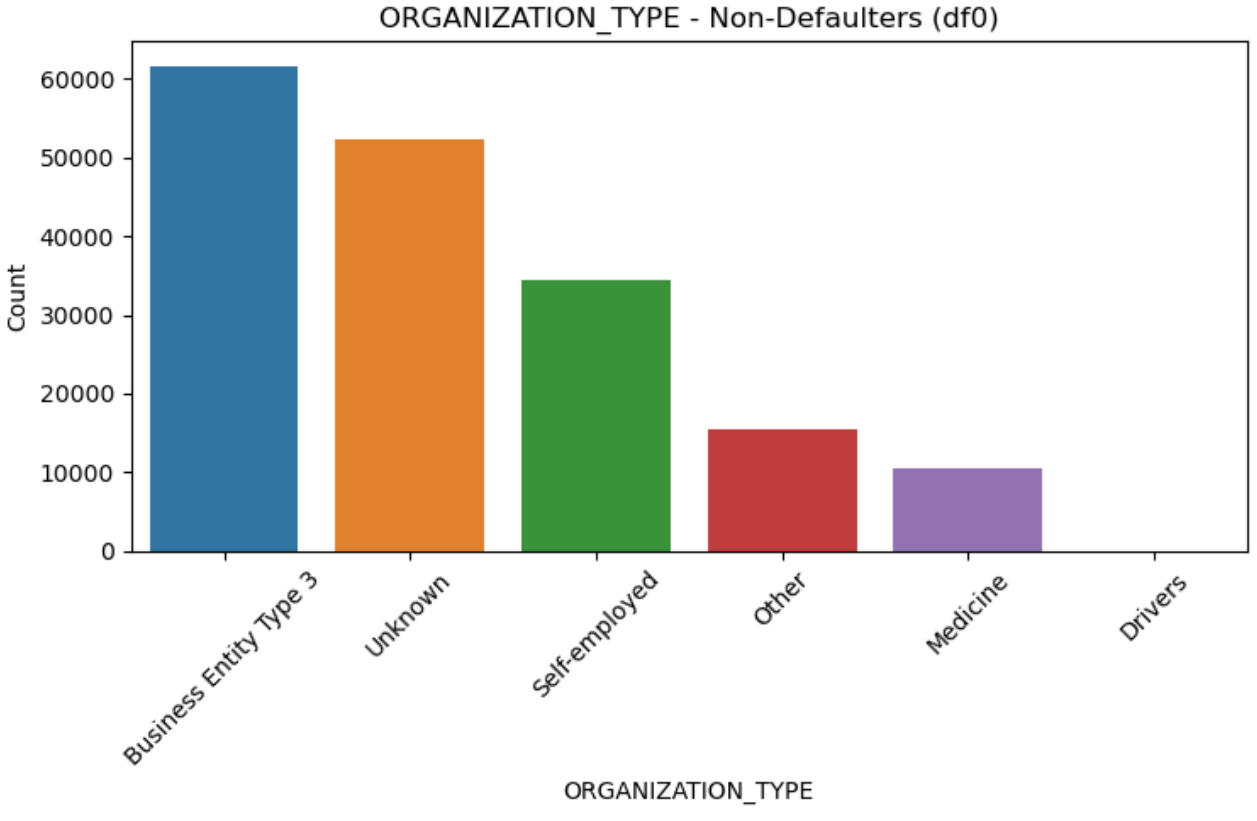


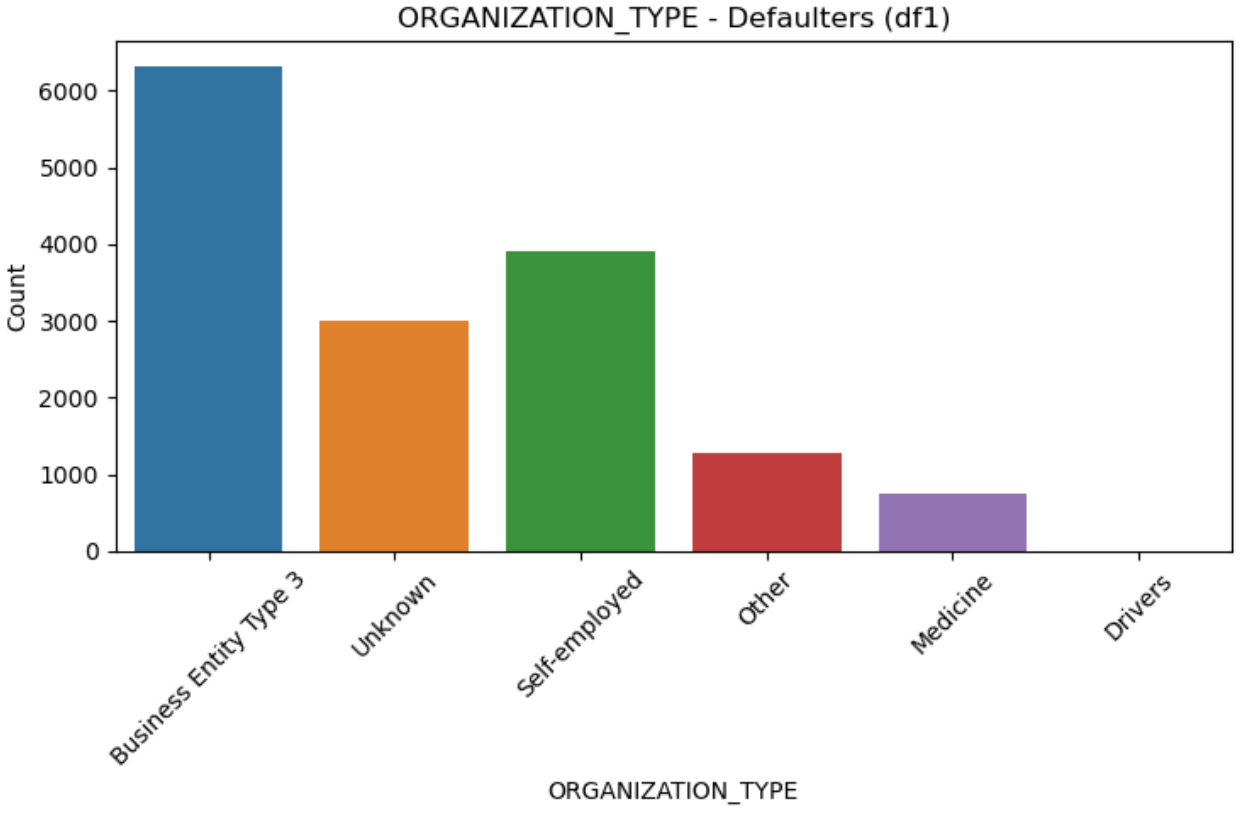


Bivariate Analysis

To cross check at columns from TARGET perspective(defaulters and non-defaulters)

- In both the TARGET groups, "**ORGANIZATION_TYPE**" column saw **"Business Entity Type 3" and "Self-Employed"** working population availed more loans than others. There were more number of rows which did not show which type of work they did, and they were grouped as **"Unknown"**

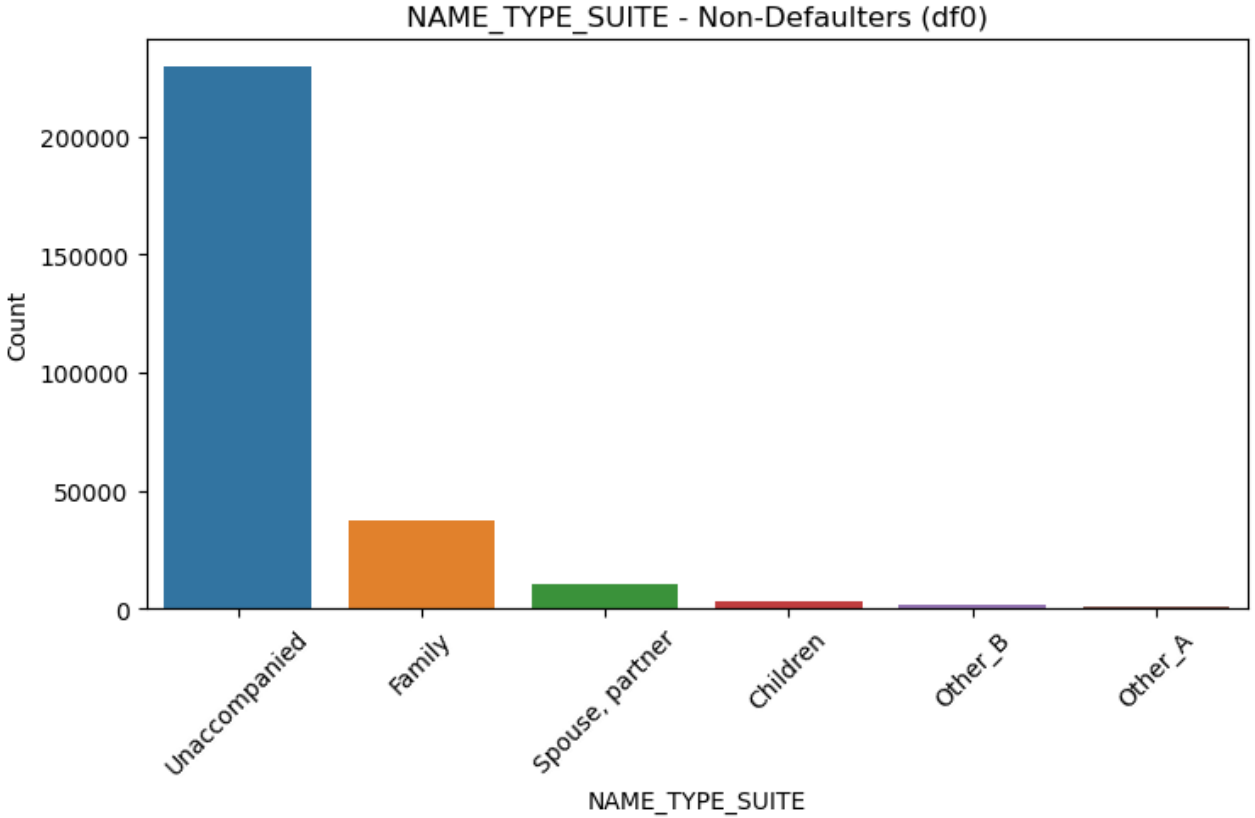


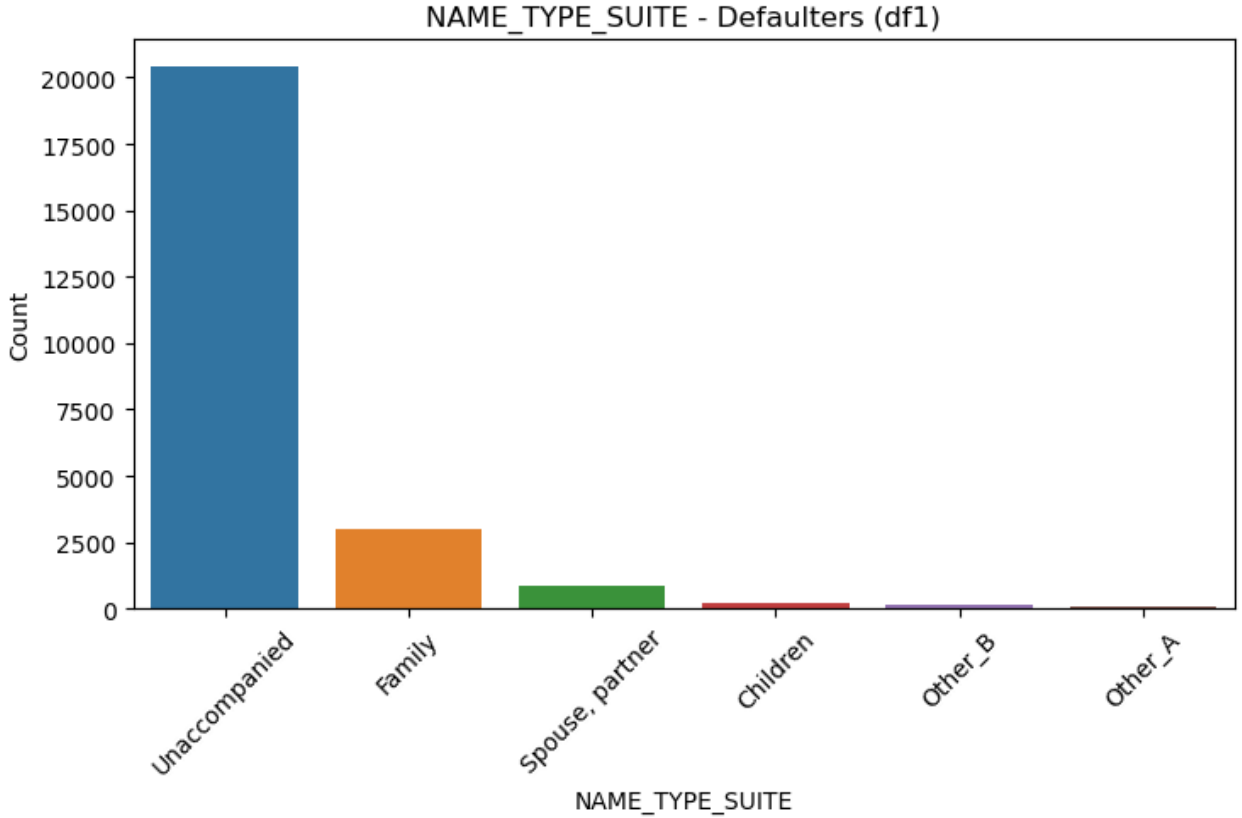


Bivariate Analysis

To cross check at columns from TARGET perspective(defaulters and non-defaulters)

- In both the TARGET groups, "**NAME_TYPE_SUITE**" column saw **"Unaccompanied"** category significantly higher. That is most of the population were not accompanied by anybody.

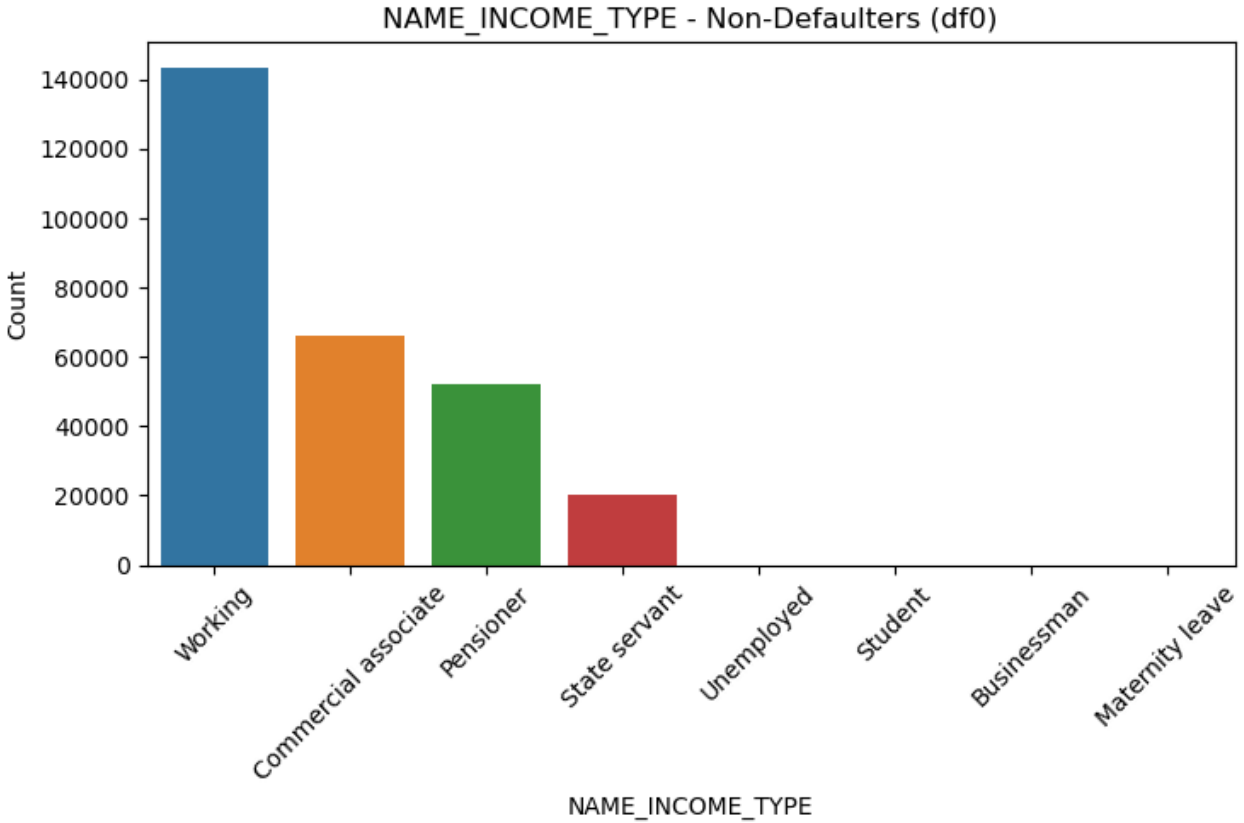


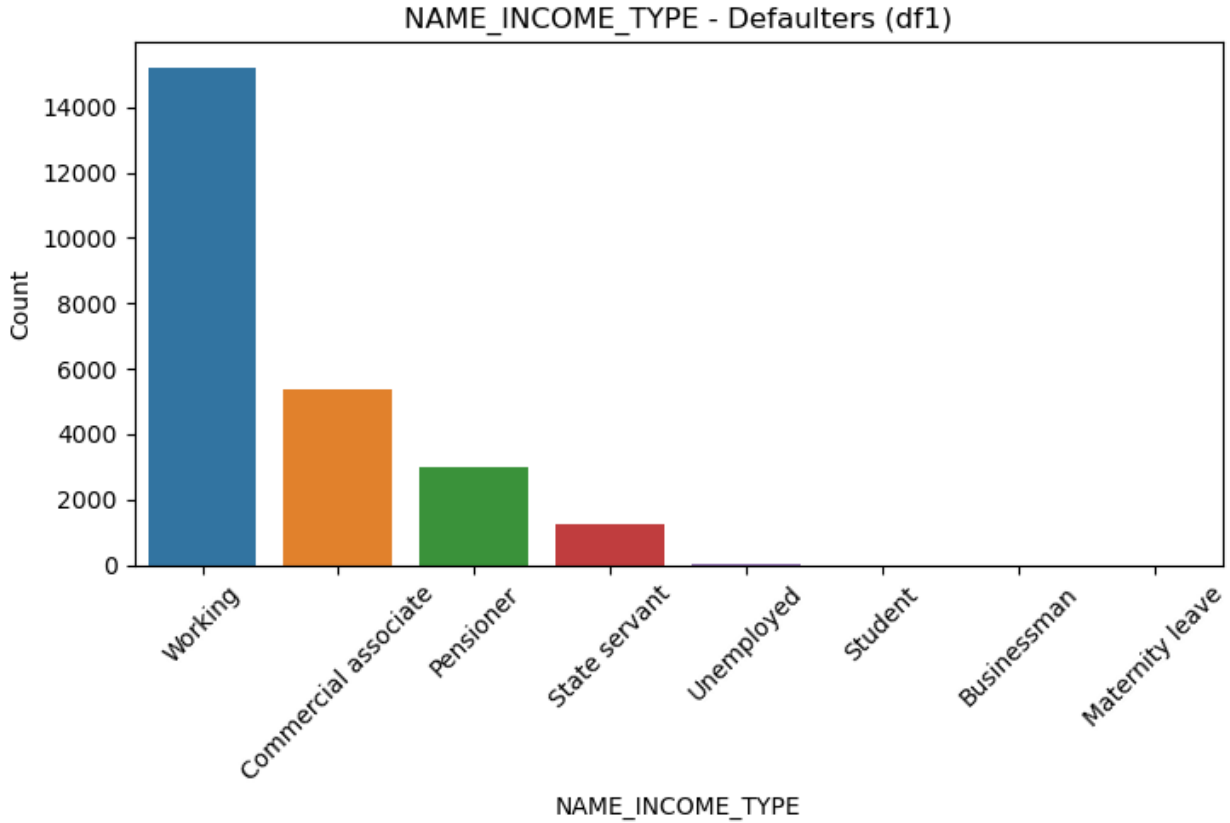


Bivariate Analysis

To cross check at columns from TARGET perspective(defaulters and non-defaulters)

- In both the TARGET groups, **working** category saw significantly large number of population followed by **Commercial associate and Pensioner** in the **NAME_INCOME_TYPE** column.

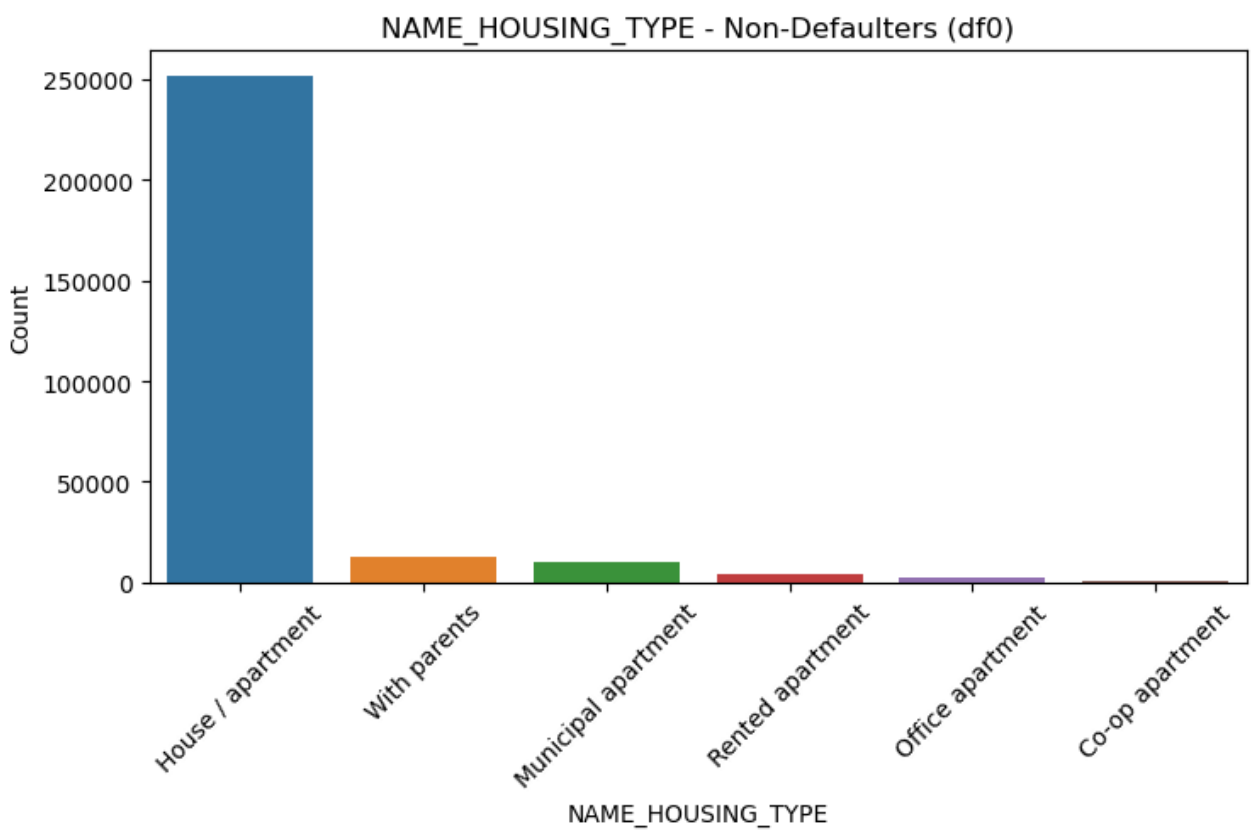


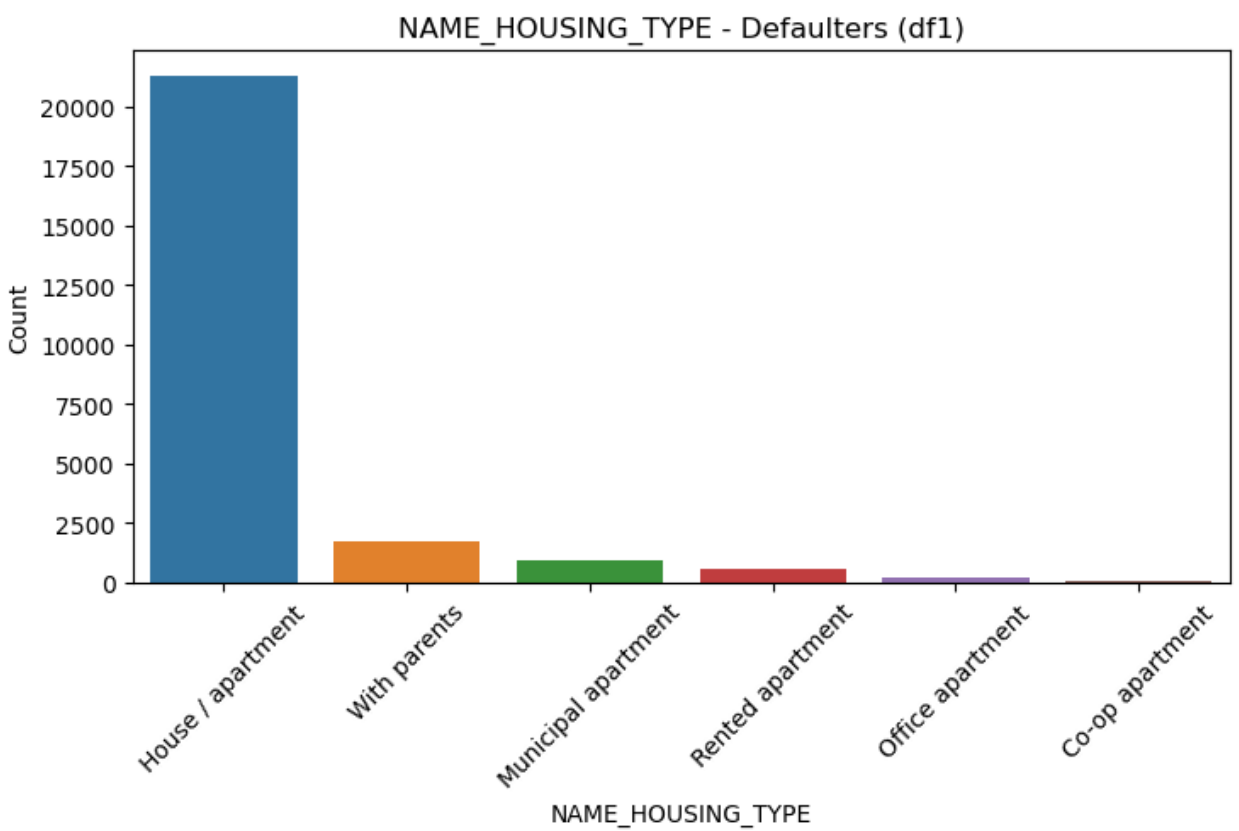


Bivariate Analysis

To cross check at columns from TARGET perspective(defaulters and non-defaulters)

- In both the TARGET groups, population living in **House/apartment** category was significantly large.

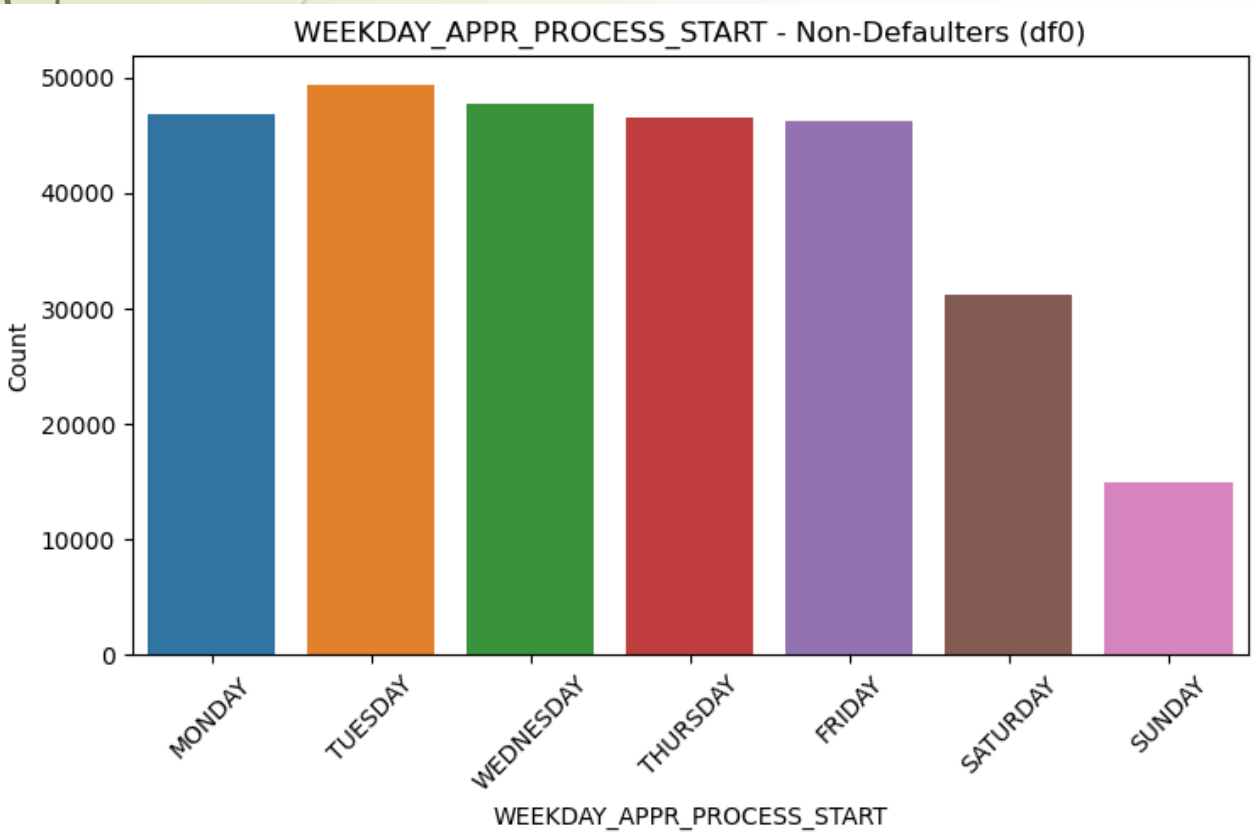


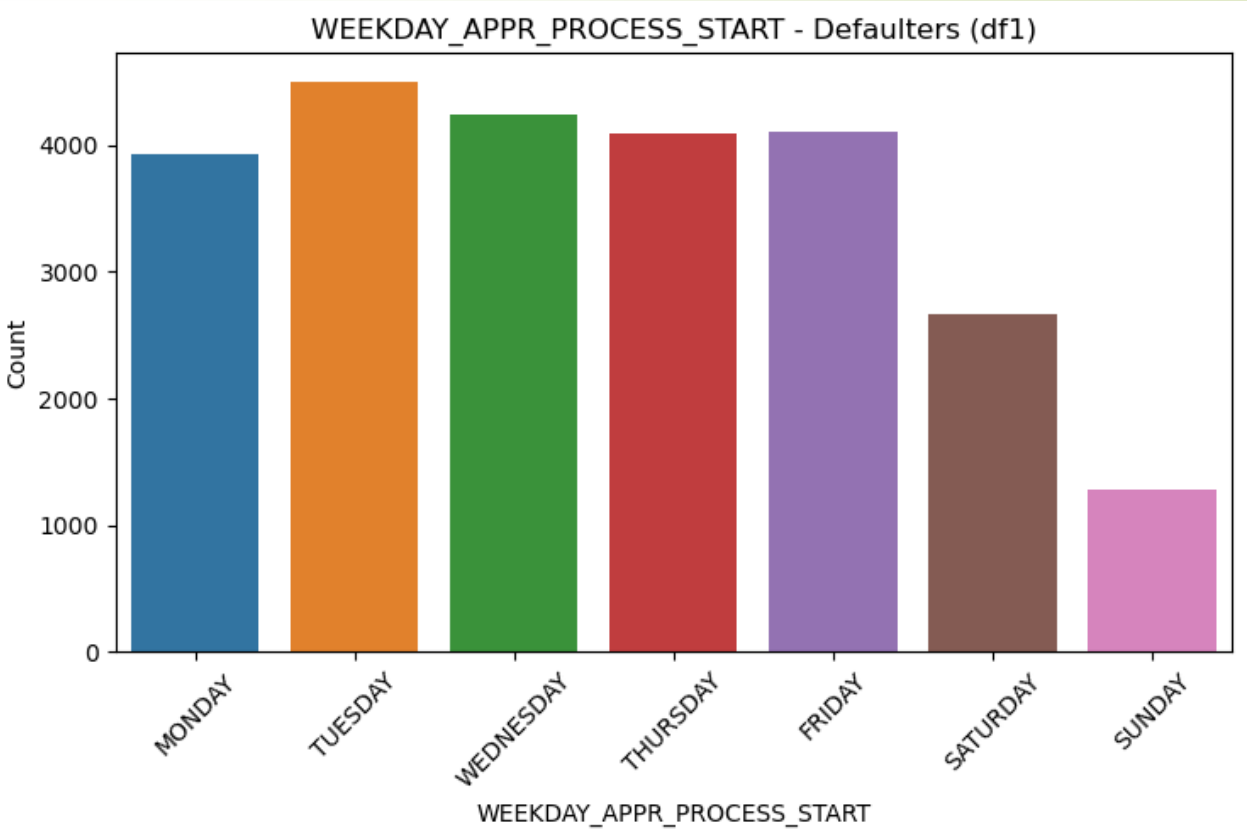


Bivariate Analysis

To cross check at columns from TARGET perspective(defaulters and non-defaulters)

- **Tuesdays** saw a increase in "**WEEKDAY_APPR_PROCESS_START**" then other days.





Multivariate Analysis

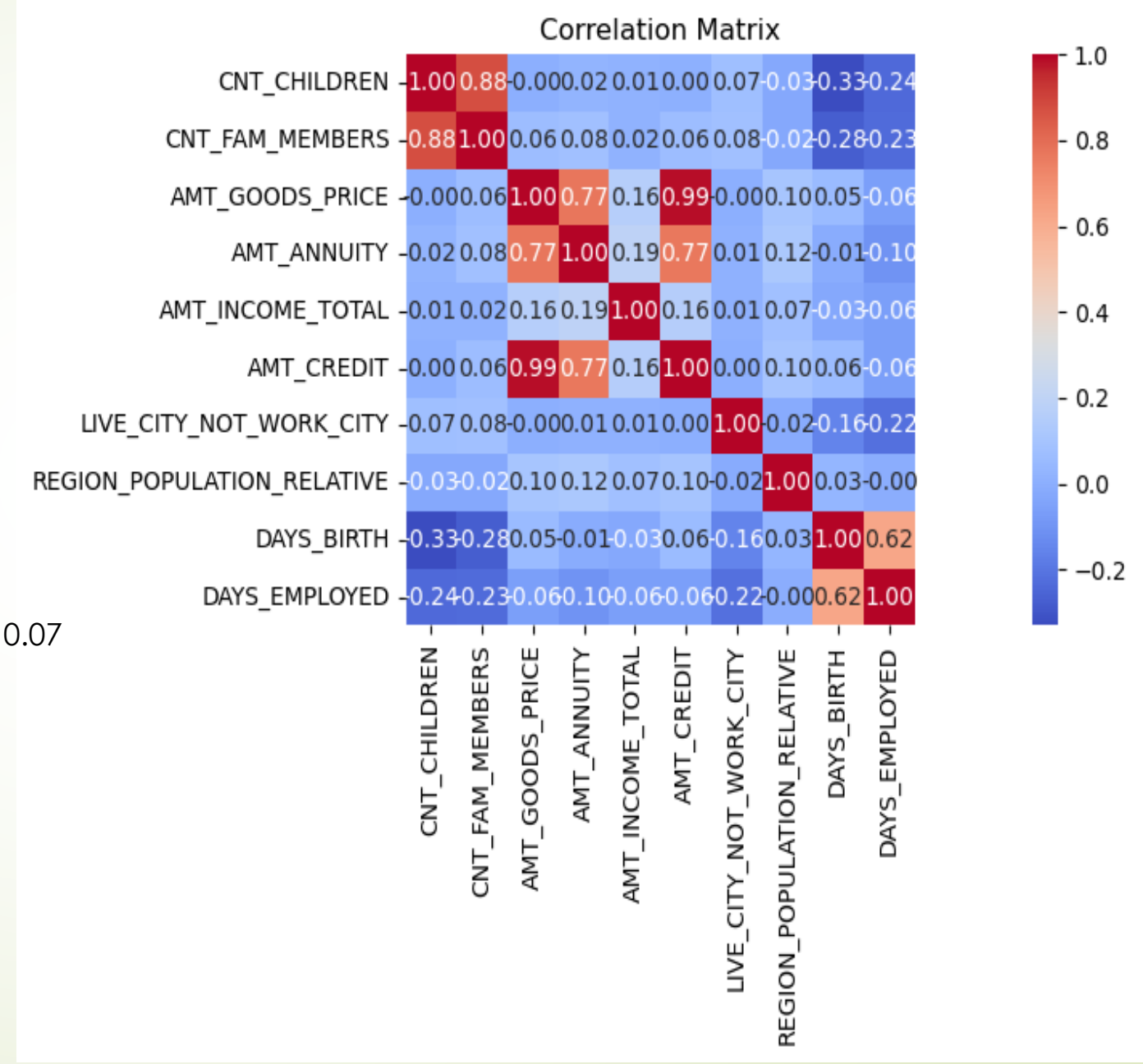
On the Overall Application Data

There is high correlation between

- AMT_CREDIT & AMT_GOODS_PRICE = 0.99
- AMT_GOODS_PRICE & AMT_ANNUITY = 0.77
- DAYS_BIRTH & DAYS_EMPLOYED = 0.62
- CNT_CHILDREN & CNT_FAM_MEMBERS = 0.88

There is low Correlation between

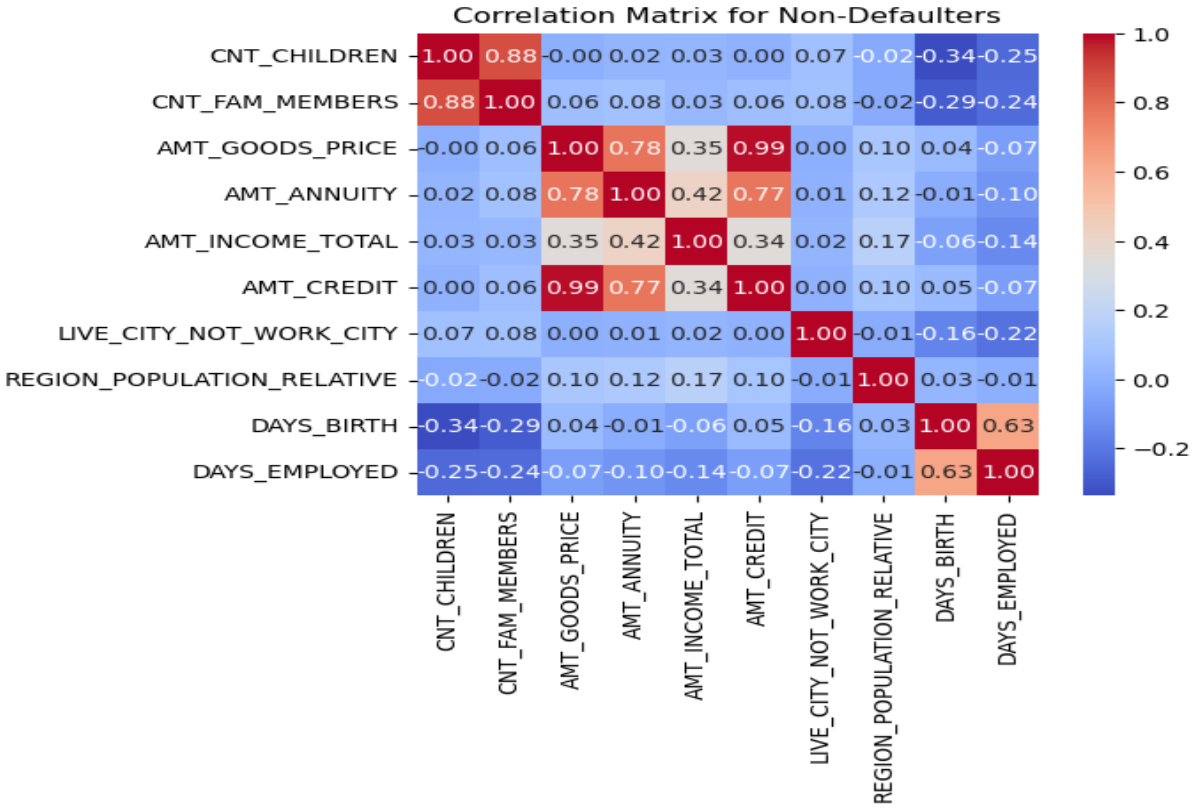
- LIVE_CITY_NOT_WORK_CITY & AMT_INCOME = 0.008
- AMT_CREDIT & CNT_FAM_MEMBERS = 0.06
- REGION_POPULATION_RELATIVE & AMT_INNCOME = 0.07

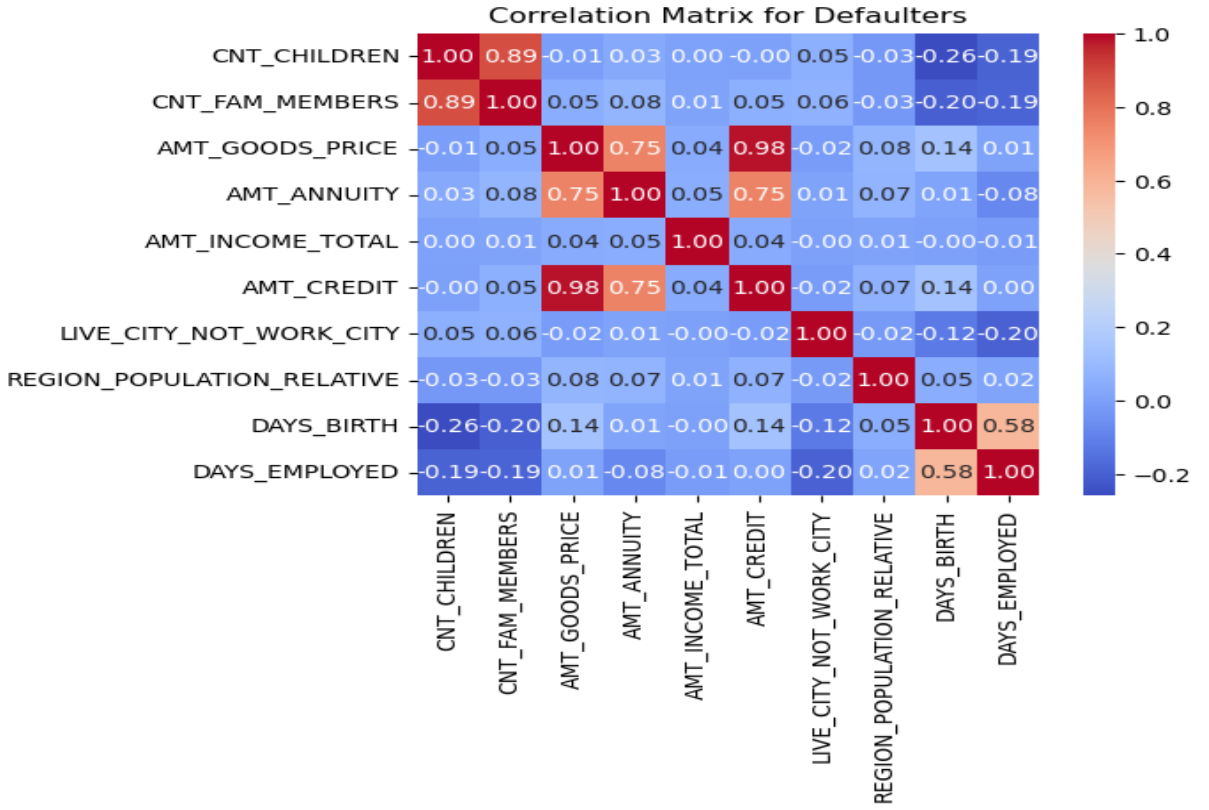


Multivariate Analysis

To cross check using TARGET columns for defaulters and non-defaulters

Among , **non-defaulters** we see **positive good correlation between AMT_INCOME_TOTAL & AMT_GOODS_PRICE , AMT_INCOME_TOTAL & AMT_ANNUITY, AMT_CREDIT & AMOUNT_INCOME_TOTAL** which is **not seen in defaulters.**





Multivariate Analysis

Correlation matrix for Non-Defaulters:

	CNT_CHILDREN	CNT_FAM_MEMBERS	AMT_GOODS_PRICE
CNT_CHILDREN	1.000000	0.878571	-0.000559
CNT_FAM_MEMBERS	0.878571	1.000000	0.062763
AMT_GOODS_PRICE	-0.000559	0.062763	1.000000
AMT_ANNUITY	0.020909	0.075789	0.776421
AMT_INCOME_TOTAL	0.027397	0.034254	0.349426
AMT_CREDIT	0.003081	0.064536	0.987022
LIVE_CITY_NOT_WORK_CITY	0.070988	0.078843	0.001285
REGION_POPULATION_RELATIVE	-0.024363	-0.023425	0.103826
DAYS_BIRTH	-0.336966	-0.285823	0.044650
DAYS_EMPLOYED	-0.245174	-0.238300	-0.068527

	AMT_ANNUITY	AMT_INCOME_TOTAL	AMT_CREDIT
CNT_CHILDREN	0.020909	0.027397	0.003081
CNT_FAM_MEMBERS	0.075789	0.034254	0.064536
AMT_GOODS_PRICE	0.776421	0.349426	0.987022
AMT_ANNUITY	1.000000	0.418948	0.771297
AMT_INCOME_TOTAL	0.418948	1.000000	0.342799
AMT_CREDIT	0.771297	0.342799	1.000000
LIVE_CITY_NOT_WORK_CITY	0.010577	0.020684	0.002506
REGION_POPULATION_RELATIVE	0.120977	0.167851	0.100604
DAYS_BIRTH	-0.012260	-0.062609	0.047378
DAYS_EMPLOYED	-0.104975	-0.140392	-0.070104

Correlation matrix for Defaulters:

	CNT_CHILDREN	CNT_FAM_MEMBERS	AMT_GOODS_PRICE
CNT_CHILDREN	1.000000	0.885484	-0.008111
CNT_FAM_MEMBERS	0.885484	1.000000	0.047367
AMT_GOODS_PRICE	-0.008111	0.047367	1.000000
AMT_ANNUITY	0.031257	0.075711	0.752295
AMT_INCOME_TOTAL	0.004796	0.006654	0.037591
AMT_CREDIT	-0.001675	0.051224	0.982783
LIVE_CITY_NOT_WORK_CITY	0.053515	0.061316	-0.016703
REGION_POPULATION_RELATIVE	-0.031975	-0.030163	0.076053
DAYS_BIRTH	-0.259109	-0.203267	0.135738
DAYS_EMPLOYED	-0.192864	-0.186515	0.006648

	AMT_ANNUITY	AMT_INCOME_TOTAL	AMT_CREDIT
CNT_CHILDREN	0.031257	0.004796	-0.001675
CNT_FAM_MEMBERS	0.075711	0.006654	0.051224
AMT_GOODS_PRICE	0.752295	0.037591	0.982783
AMT_ANNUITY	1.000000	0.046421	0.752195
AMT_INCOME_TOTAL	0.046421	1.000000	0.038131
AMT_CREDIT	0.752195	0.038131	1.000000
LIVE_CITY_NOT_WORK_CITY	0.009902	-0.001353	-0.016509
REGION_POPULATION_RELATIVE	0.071690	0.009135	0.069161
DAYS_BIRTH	0.014303	-0.003096	0.135316
DAYS_EMPLOYED	-0.081207	-0.014977	0.001930

Multivariate Analysis

Non-Defaulters

	LIVE_CITY_NOT_WORK_CITY
CNT_CHILDREN	0.070988
CNT_FAM_MEMBERS	0.078843
AMT_GOODS_PRICE	0.001285
AMT_ANNUITY	0.010577
AMT_INCOME_TOTAL	0.020684
AMT_CREDIT	0.002506
LIVE_CITY_NOT_WORK_CITY	1.000000
REGION_POPULATION_RELATIVE	-0.013502
DAYS_BIRTH	-0.160072
DAYS_EMPLOYED	-0.221387

	REGION_POPULATION_RELATIVE	DAYS_BIRTH
CNT_CHILDREN	-0.024363	-0.336966
CNT_FAM_MEMBERS	-0.023425	-0.285823
AMT_GOODS_PRICE	0.103826	0.044650
AMT_ANNUITY	0.120977	-0.012260
AMT_INCOME_TOTAL	0.167851	-0.062609
AMT_CREDIT	0.100604	0.047378
LIVE_CITY_NOT_WORK_CITY	-0.013502	-0.160072
REGION_POPULATION_RELATIVE	1.000000	0.025244
DAYS_BIRTH	0.025244	1.000000
DAYS_EMPLOYED	-0.007198	0.626114

Defaulters

	LIVE_CITY_NOT_WORK_CITY
CNT_CHILDREN	0.053515
CNT_FAM_MEMBERS	0.061316
AMT_GOODS_PRICE	-0.016703
AMT_ANNUITY	0.009902
AMT_INCOME_TOTAL	-0.001353
AMT_CREDIT	-0.016509
LIVE_CITY_NOT_WORK_CITY	1.000000
REGION_POPULATION_RELATIVE	-0.020428
DAYS_BIRTH	-0.123623
DAYS_EMPLOYED	-0.198484

	REGION_POPULATION_RELATIVE	DAYS_BIRTH
CNT_CHILDREN	-0.031975	-0.259109
CNT_FAM_MEMBERS	-0.030163	-0.203267
AMT_GOODS_PRICE	0.076053	0.135738
AMT_ANNUITY	0.071690	0.014303
AMT_INCOME_TOTAL	0.009135	-0.003096
AMT_CREDIT	0.069161	0.135316
LIVE_CITY_NOT_WORK_CITY	-0.020428	-0.123623
REGION_POPULATION_RELATIVE	1.000000	0.048190
DAYS_BIRTH	0.048190	1.000000
DAYS_EMPLOYED	0.015532	0.582185

Multivariate Analysis

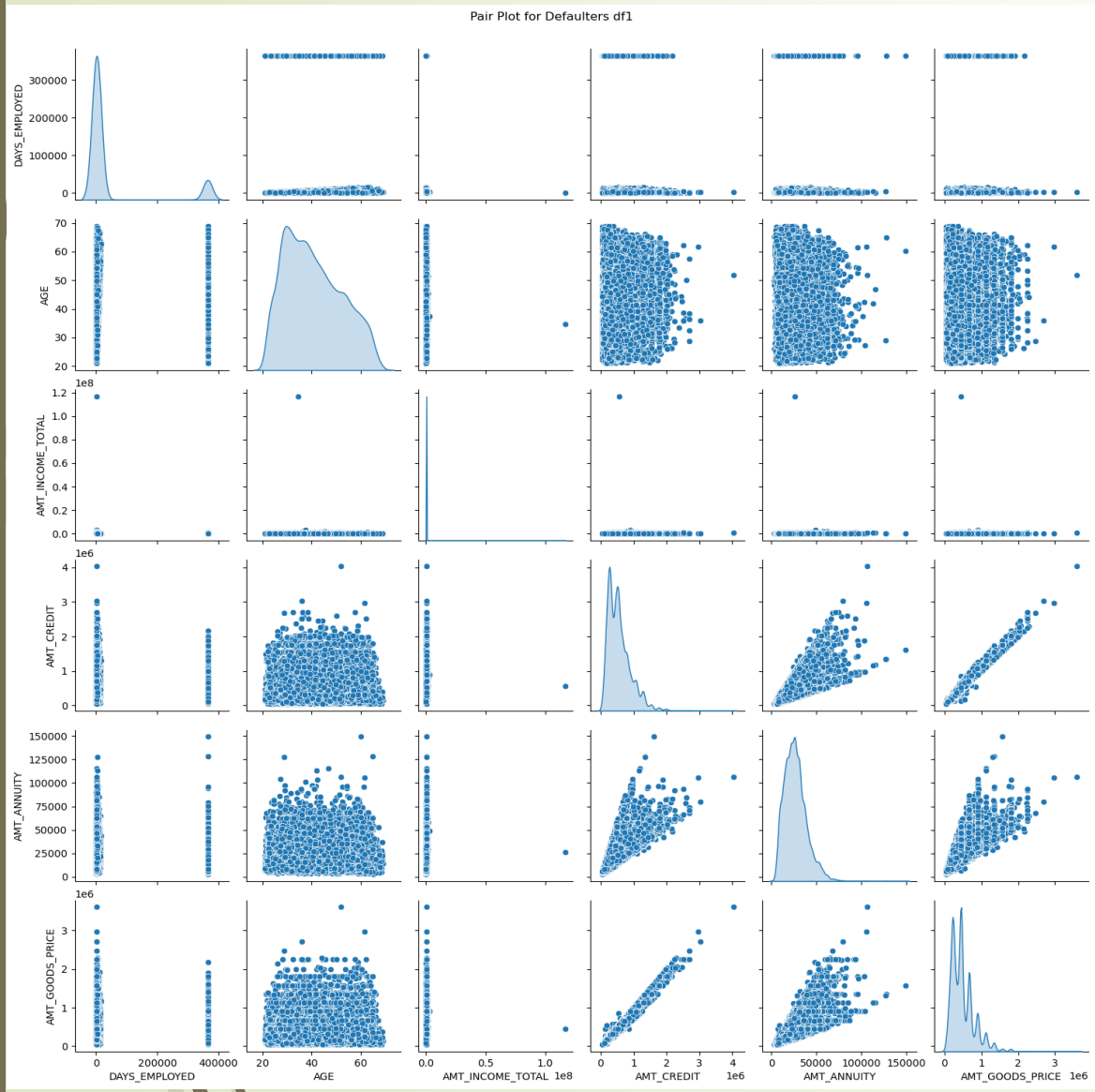
Non-Defaulters

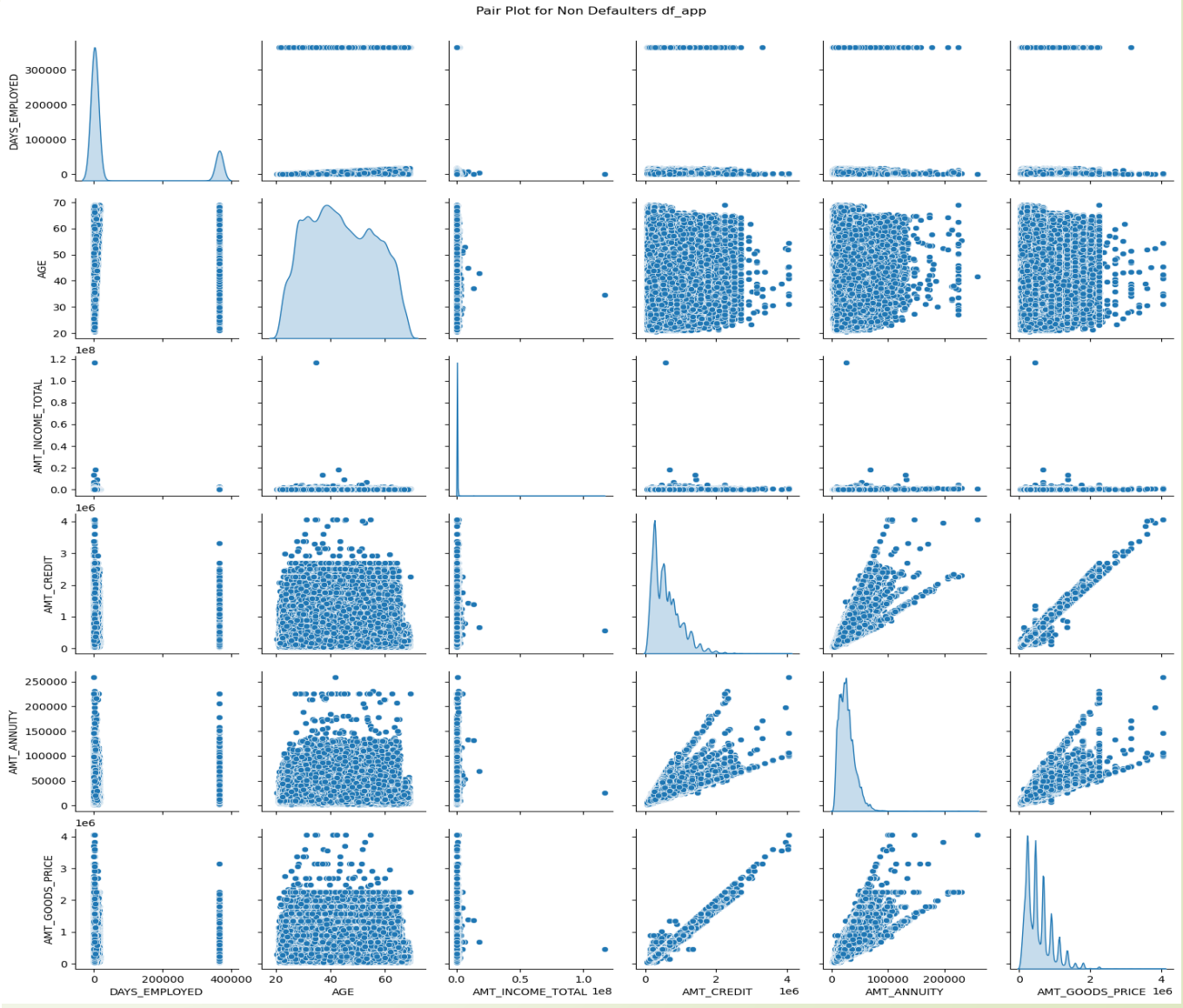
	DAYS_EMPLOYED
CNT_CHILDREN	-0.245174
CNT_FAM_MEMBERS	-0.238300
AMT_GOODS_PRICE	-0.068527
AMT_ANNUITY	-0.104975
AMT_INCOME_TOTAL	-0.140392
AMT_CREDIT	-0.070104
LIVE_CITY_NOT_WORK_CITY	-0.221387
REGION_POPULATION_RELATIVE	-0.007198
DAYS_BIRTH	0.626114
DAYS_EMPLOYED	1.000000

Defaulters

	DAYS_EMPLOYED
CNT_CHILDREN	-0.192864
CNT_FAM_MEMBERS	-0.186515
AMT_GOODS_PRICE	0.006648
AMT_ANNUITY	-0.081207
AMT_INCOME_TOTAL	-0.014977
AMT_CREDIT	0.001930
LIVE_CITY_NOT_WORK_CITY	-0.198484
REGION_POPULATION_RELATIVE	0.015532
DAYS_BIRTH	0.582185
DAYS_EMPLOYED	1.000000

Multivariate Analysis





Previous Data: Cleaning

1. 37 columns were reduced to 26 as 11 columns had null values which were more than 40%.
2. 7 columns had less than 23% null values .
    1. Used median statistics for the following columns : AMT_GOODS_PRICE, AMT_ANNUITY, AMT_CREDIT, CNT_PAYMENT, PRODUCT_COMBINATION
    2. Did not impute with anything or capped the null rows for DAYS_FIRST_DUE and DAYS_FIRST_DRAWING.

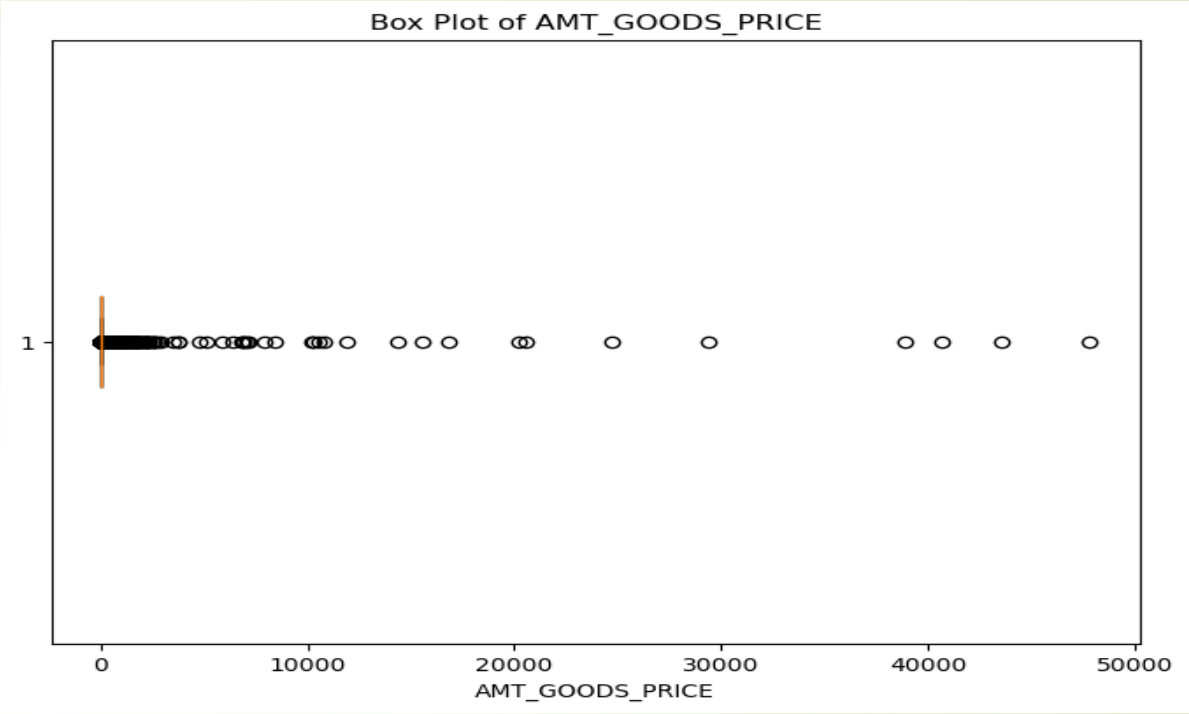
Previous Data: Anomalies

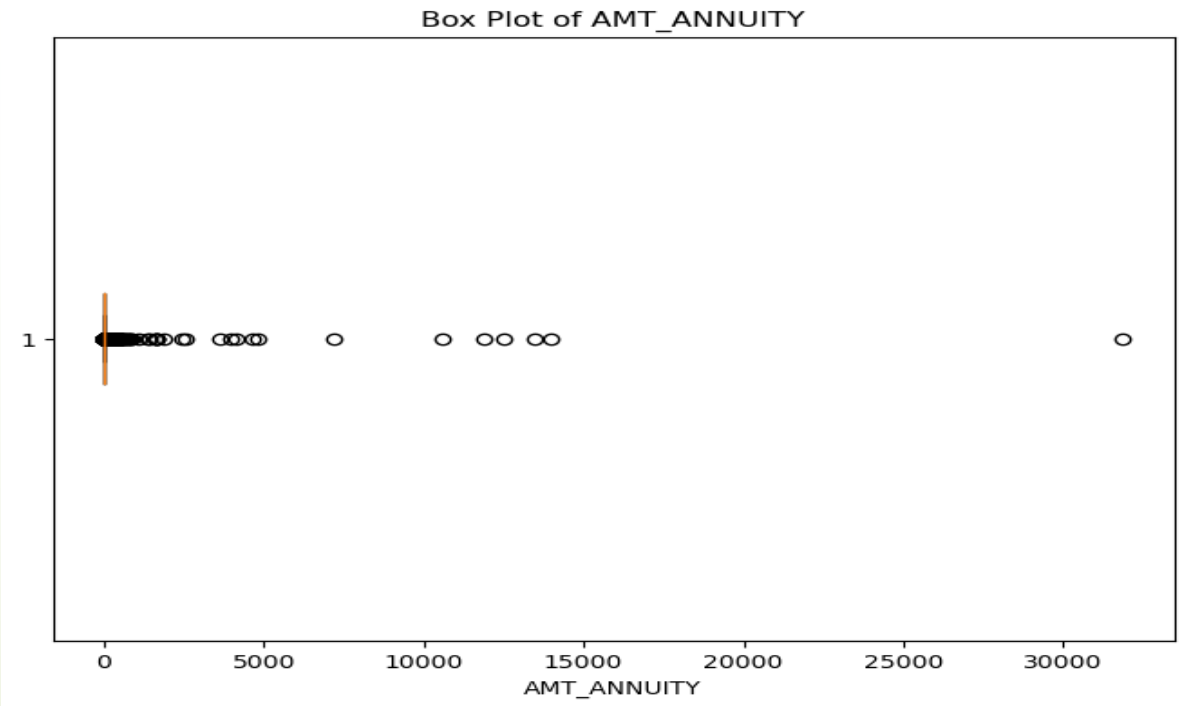
1. NAME_CONTRACT_TYPE and NAME_CLIENT_TYPE columns had “XNA” rows replaced it with “Unknown”
2. NAME_PAYMENT_TYPE column had “XNA” rows with significantly large numbers , hence did not make any changes.

Univariate Analysis

AMT_GOODS_PRICE
AMT_ANNUITY:

Both the columns have outliers and data points are beyond 75th percentile.

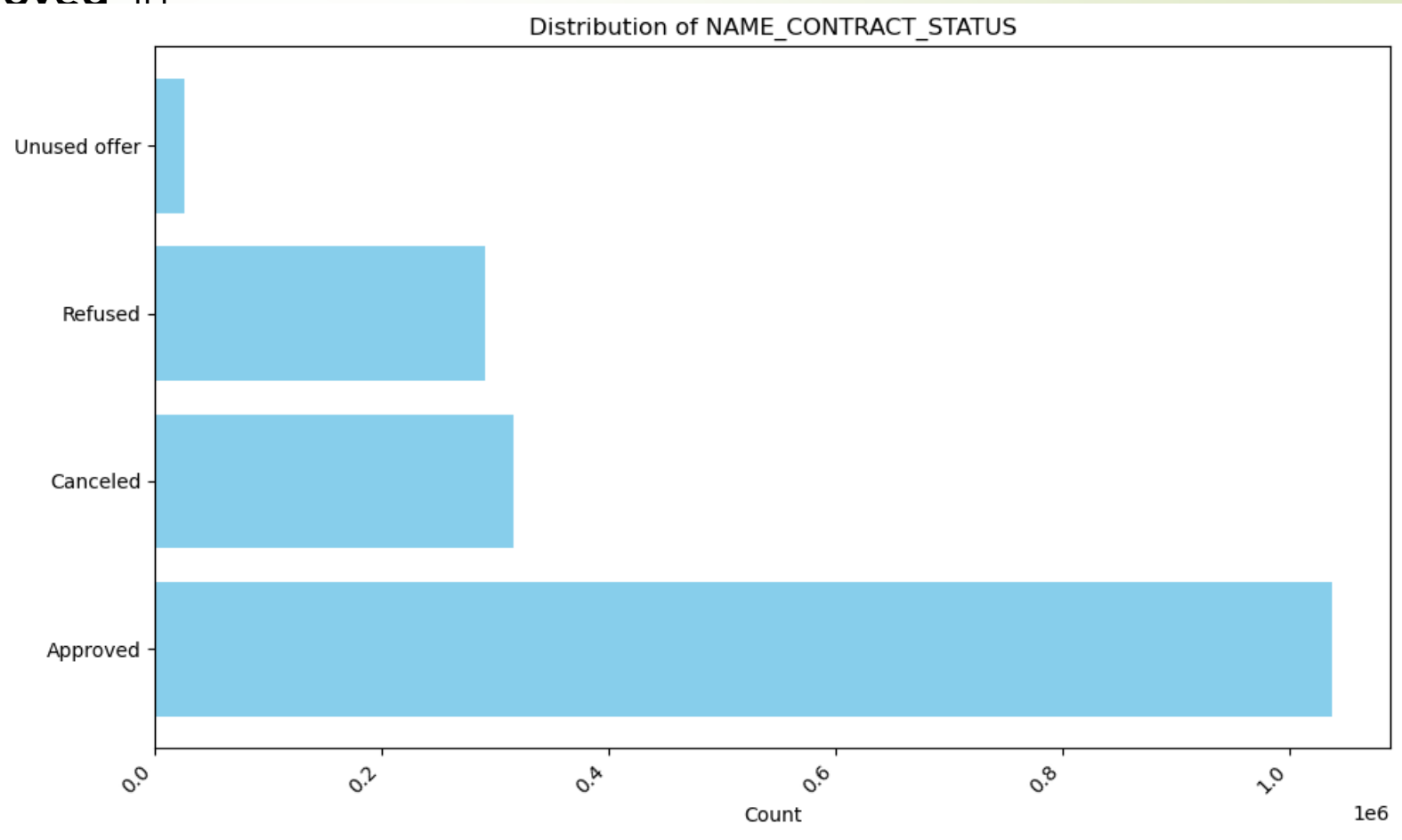




Univariate Analysis

NAME_CONTRACT_STATUS

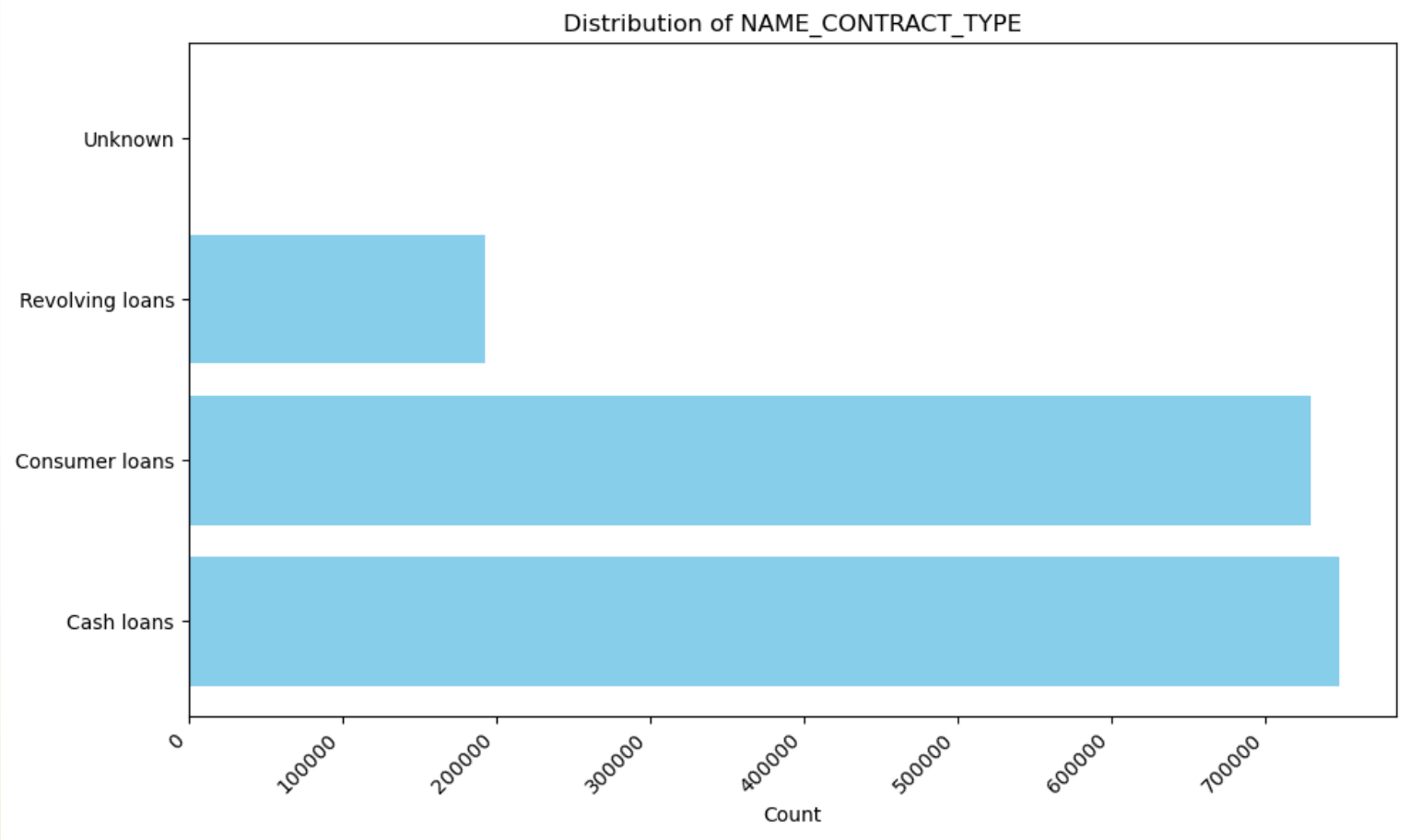
For most of the clients loan was **Approved** in the previous application data



Univariate Analysis

NAME_CONTRACT_TYPE

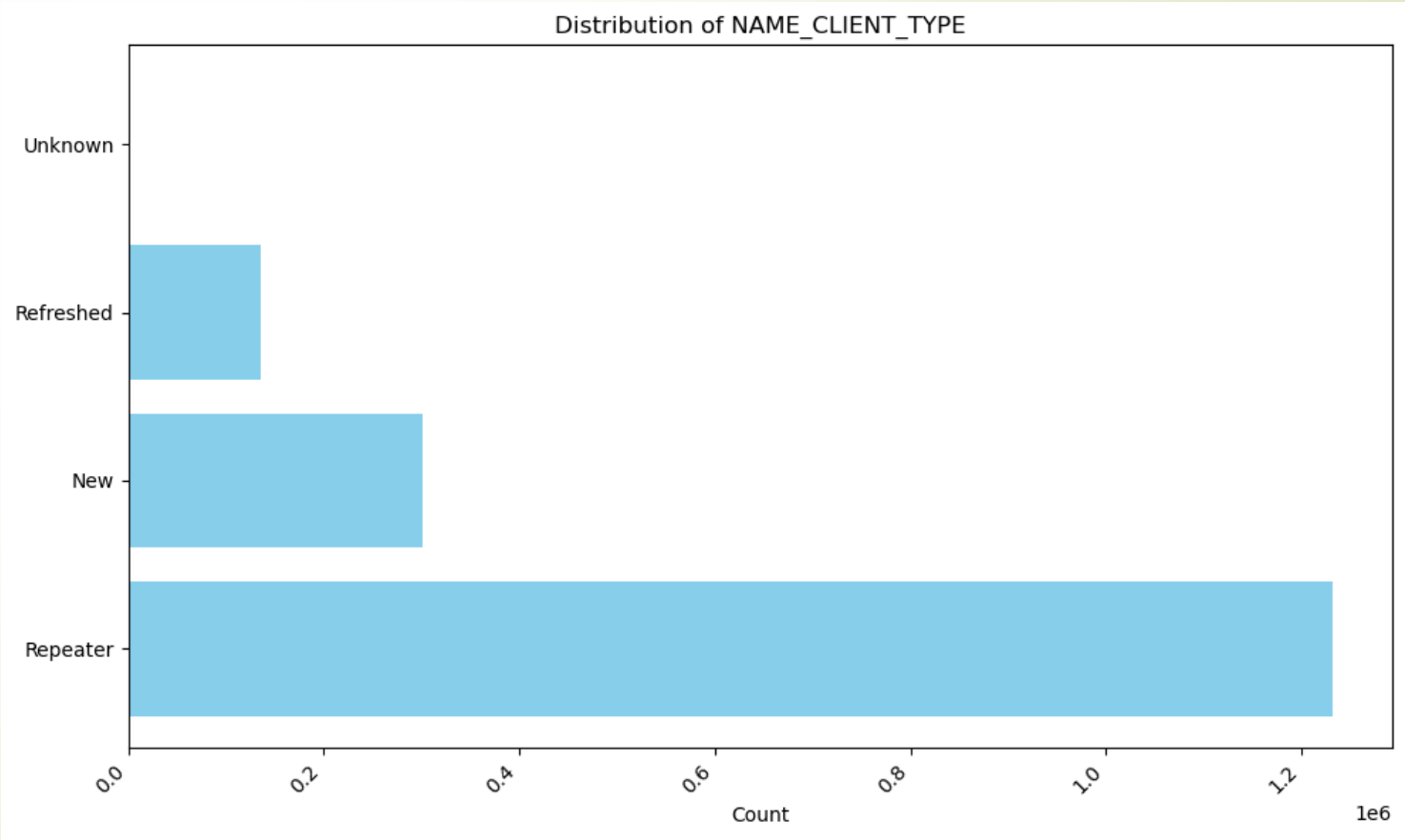
Cash and Consumer loans are the most common contract product of the previous clients.



Univariate Analysis

NAME_CLIENT_TYPE

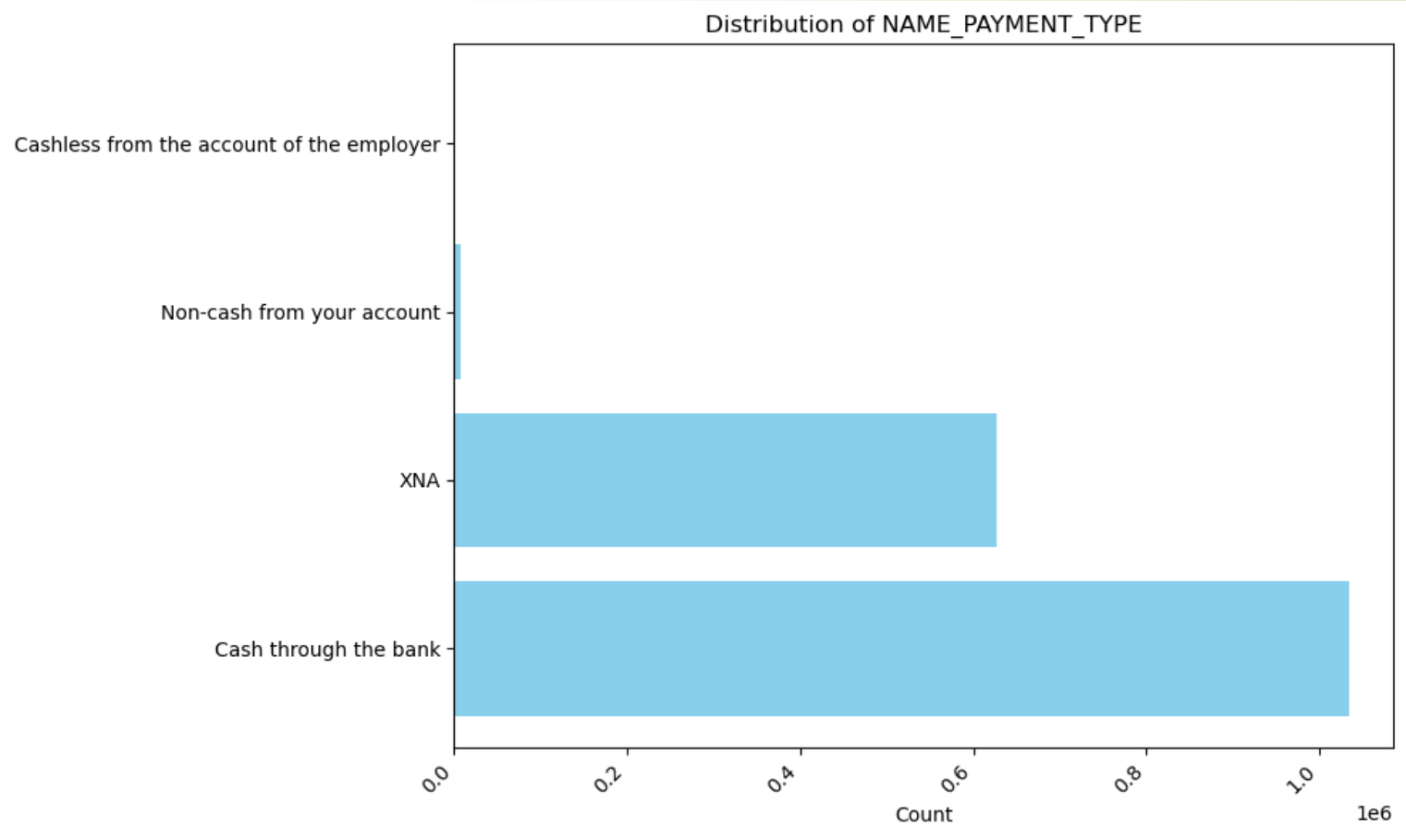
Repeaters are the most common clients seen . They have come back for more loans.



Univariate Analysis

NAME_PAYMENT_TYPE

Cash through the bank is the most preferred method used by the previous clients



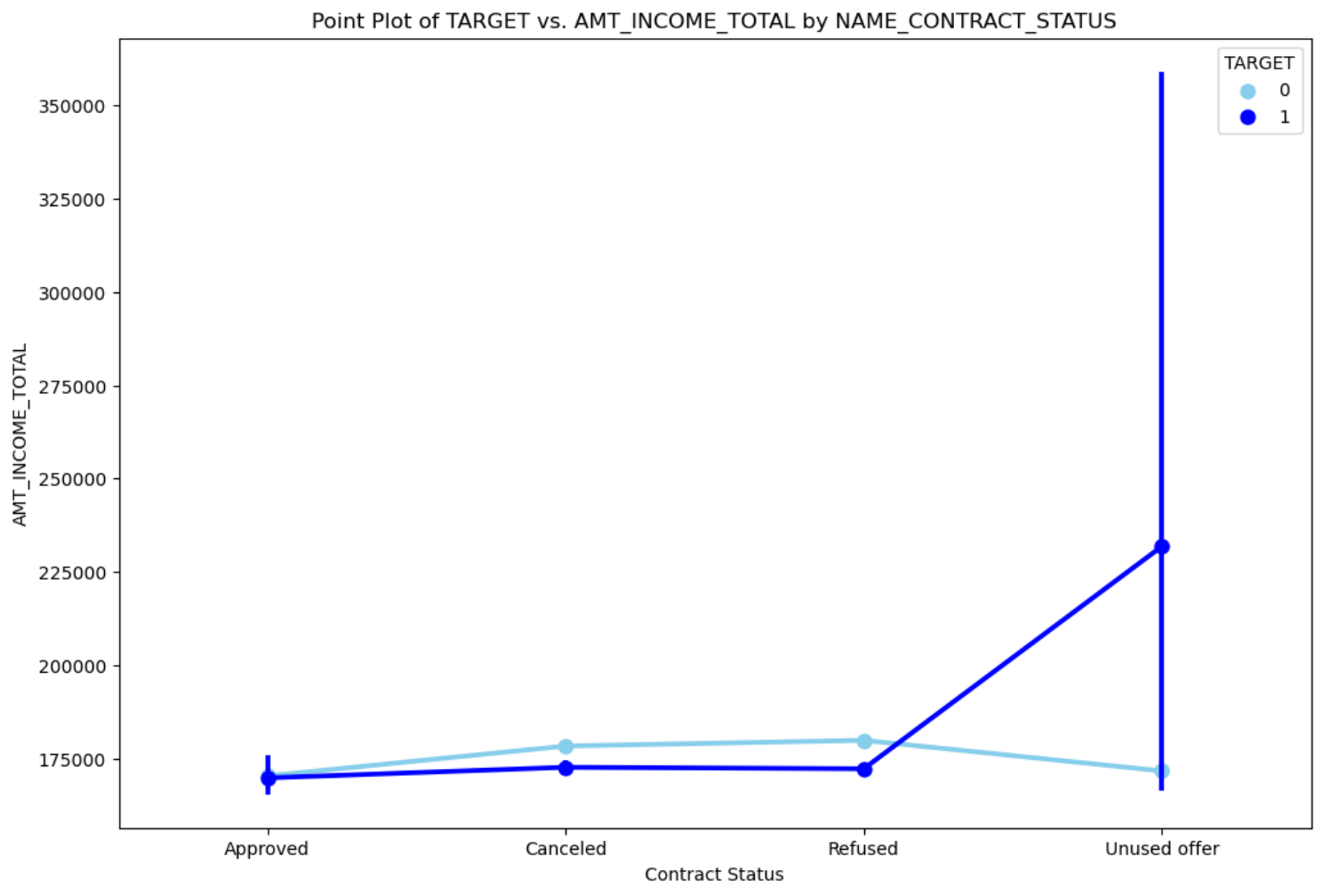
MERGED DATA

Both the data frames were merged using SK_ID_CURR.
With this we have 78 rows and 1,413,701 columns

Multivariate Analysis

TARGET column comparison with NAME_CONTRACT_STATUS & AMT_INCOME_TOTAL

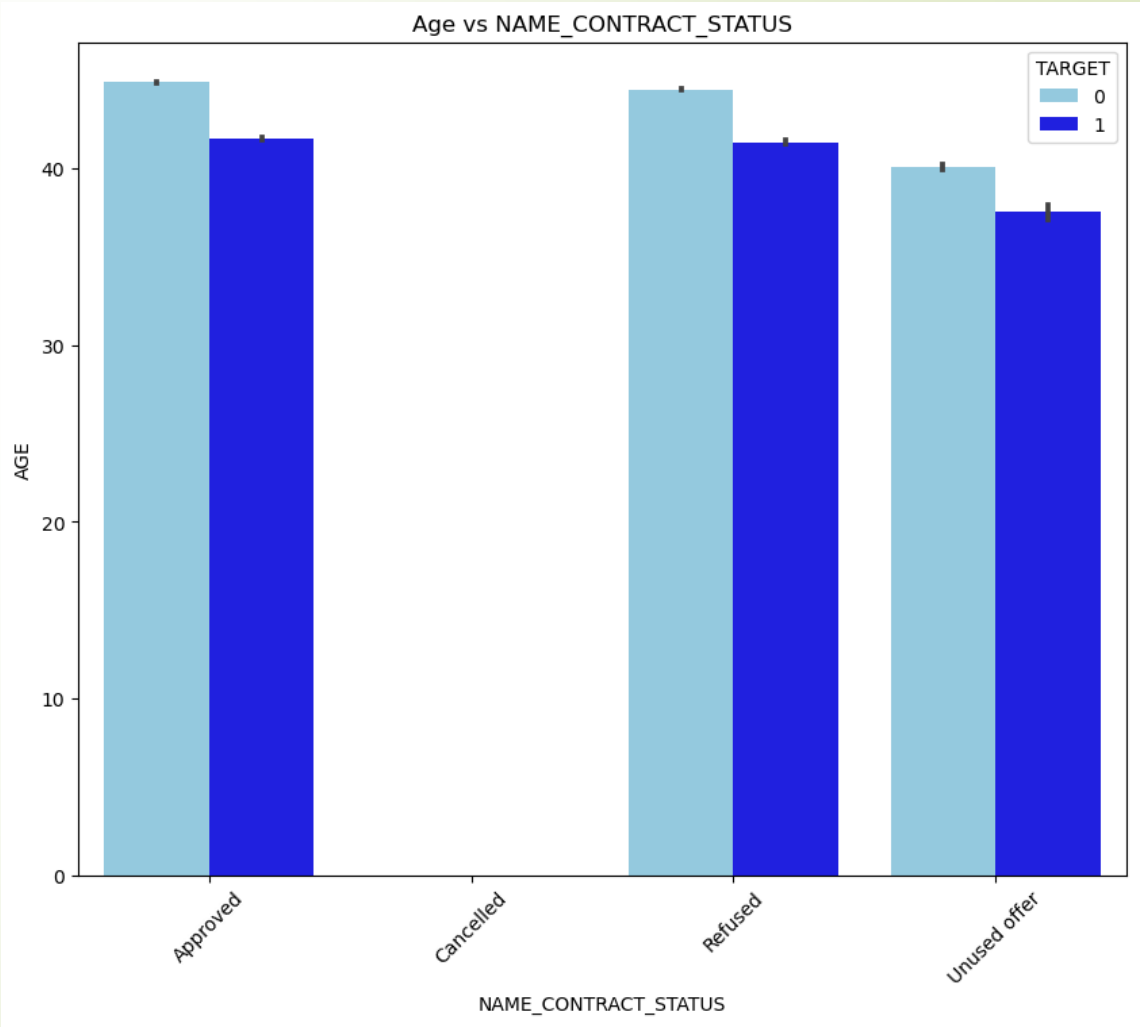
There are significant number of "**Unused offers**" by the defaulted clients who have higher income.



Multivariate Analysis

TARGET column comparison with NAME_CONTRACT_STATUS & AGE

Loan Approval rate is more than Cancelled.
Clients who have Refused or Unused the loan offer is more among both defaulters and non-defaulters.

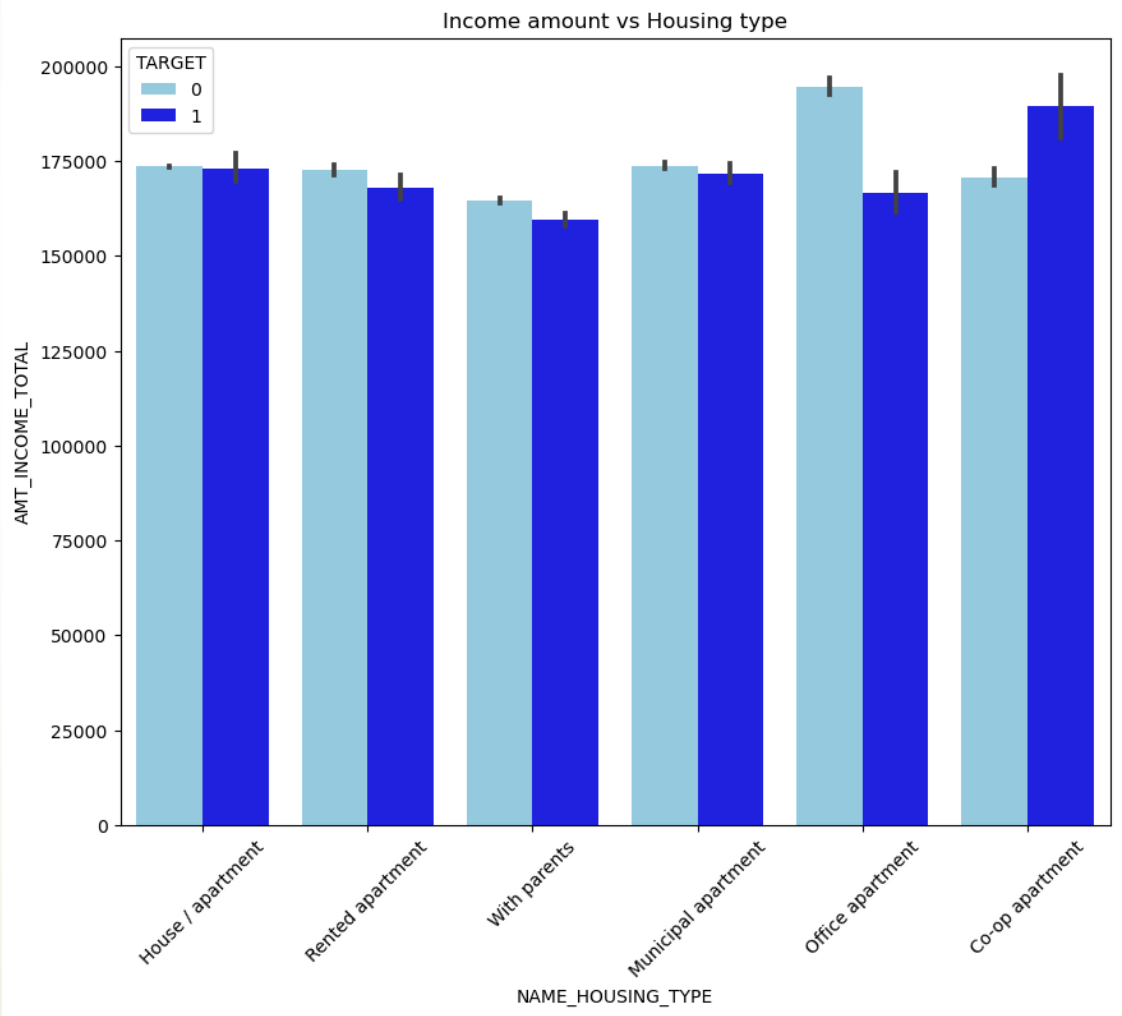


Multivariate Analysis

TARGET column comparison with NAME_HOUSING_TYPE & AMT_INCOME_TOTAL

There are significant number of "**Office apartment**" housing type by **the non-defaulted** clients who have **higher income**.

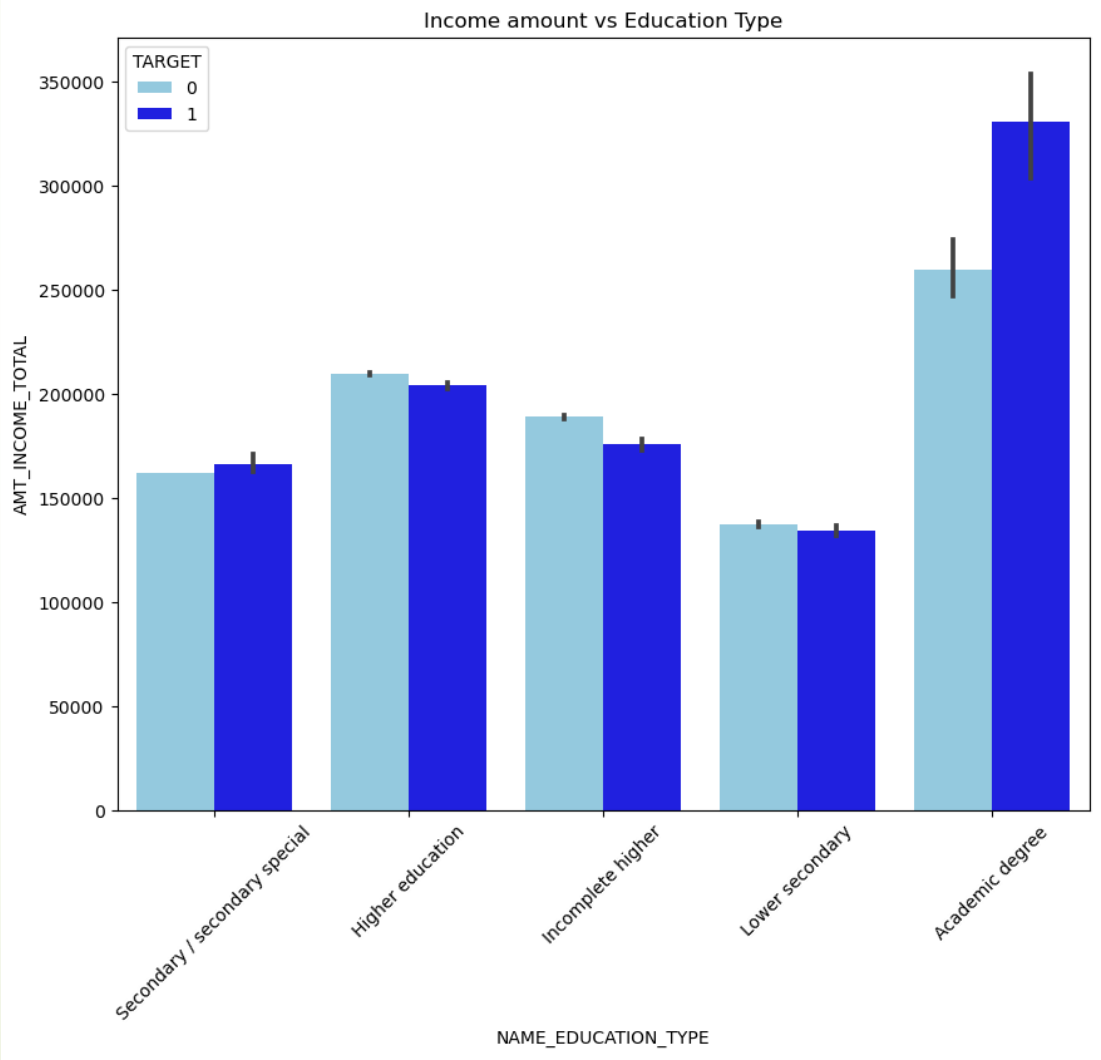
Client who lived in "**Co-op apartment**" housing type by the **defaulters** also had **higher income**.



Multivariate Analysis

TARGET column comparison with EDUCATION TYPE & AMT_INCOME_TOTAL

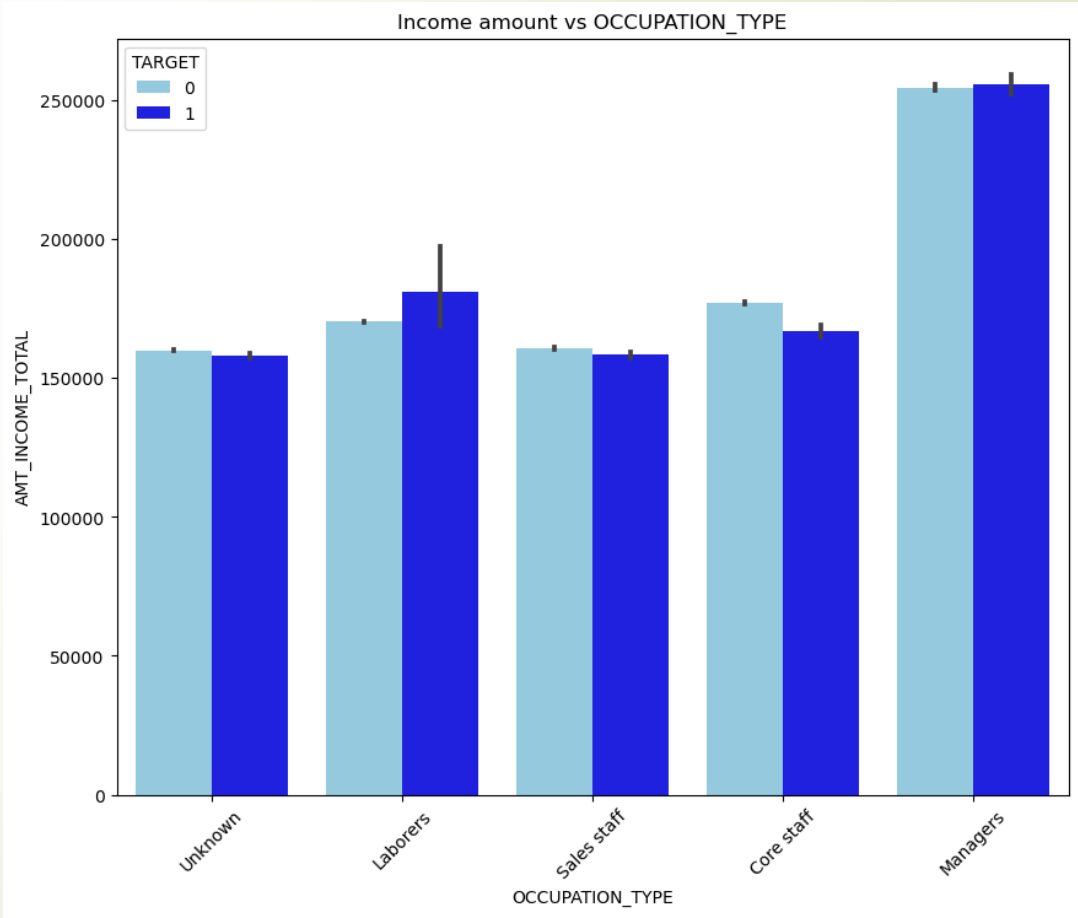
There are significantly higher number of "**Academic degree"** holders by **both defaulter and non-defaulter** clients who have **higher income**.



Multivariate Analysis

TARGET column comparison with OCCUPATION_TYPE & AMT_INCOME_TOTAL

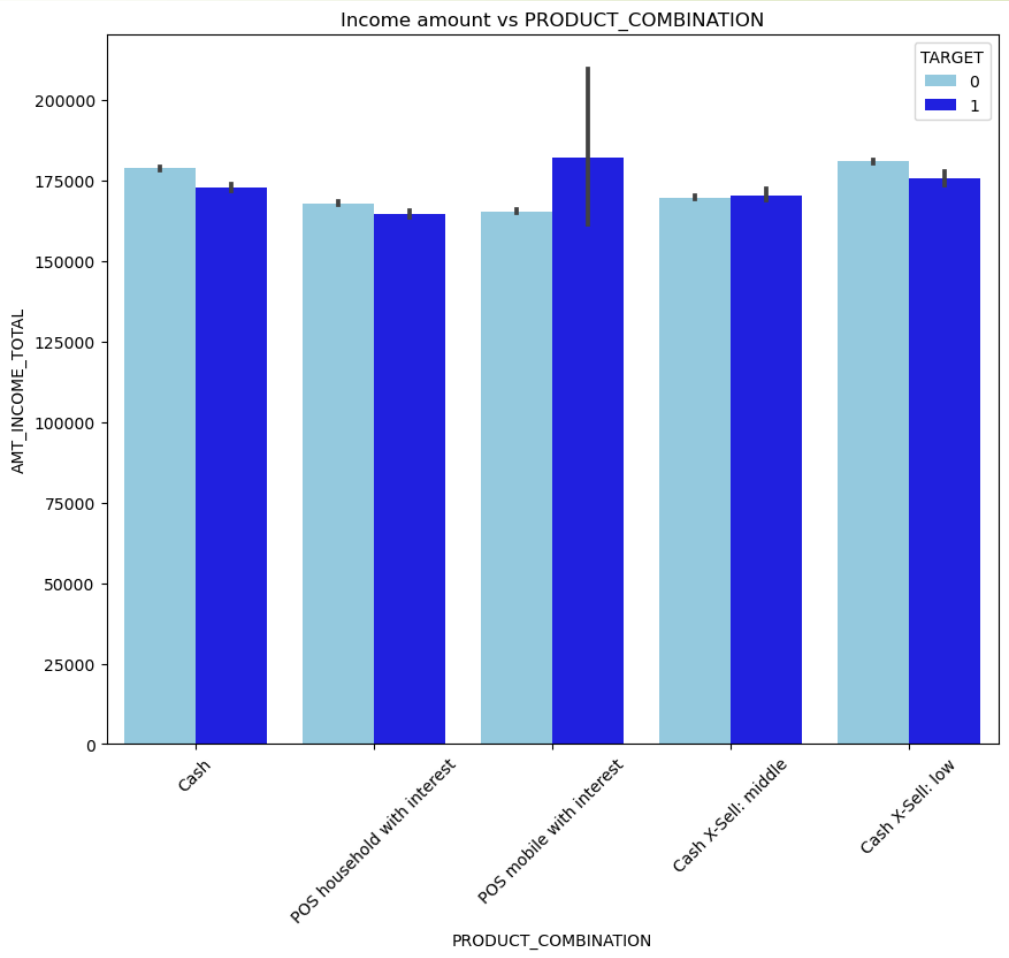
There are significantly higher number of "**Managers**" in **both defaulter and non-defaulter** clients who have **higher income**.



Multivariate Analysis

TARGET column comparison with PRODUCT_COMBINATION & AMT_INCOME_TOTAL

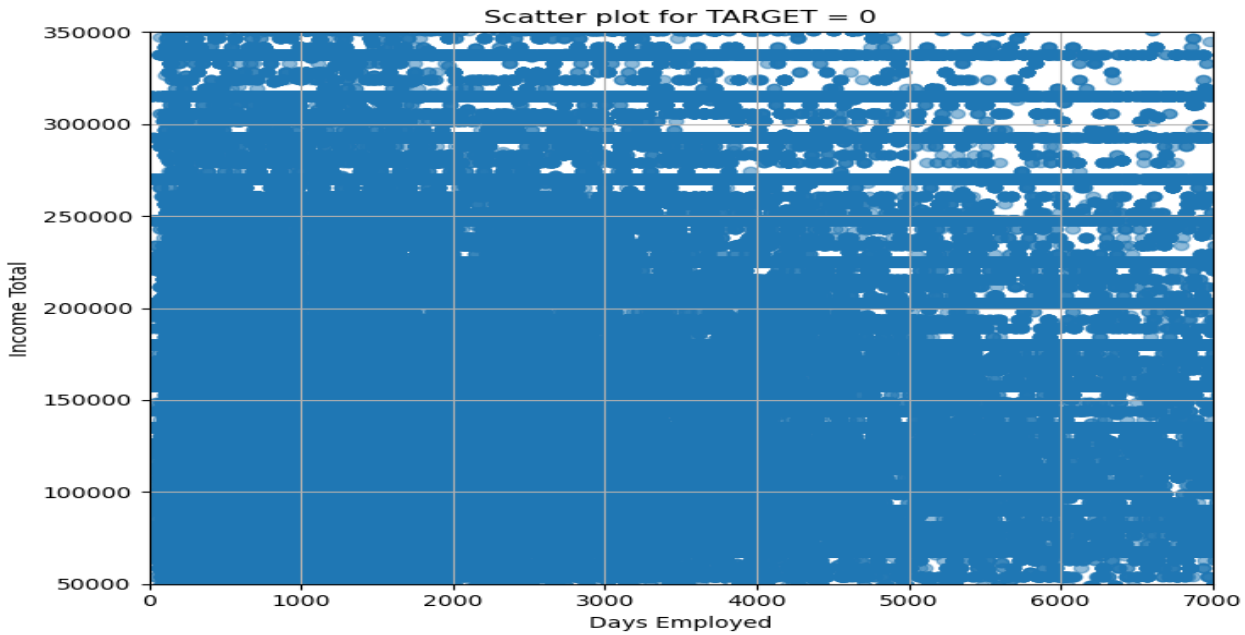
Population who purchased "**POS Mobile with interest" in the Defaulter** clients have higher income and population who bought "**Cash x-sell low"** among the **Non-defaulters** have higher income.

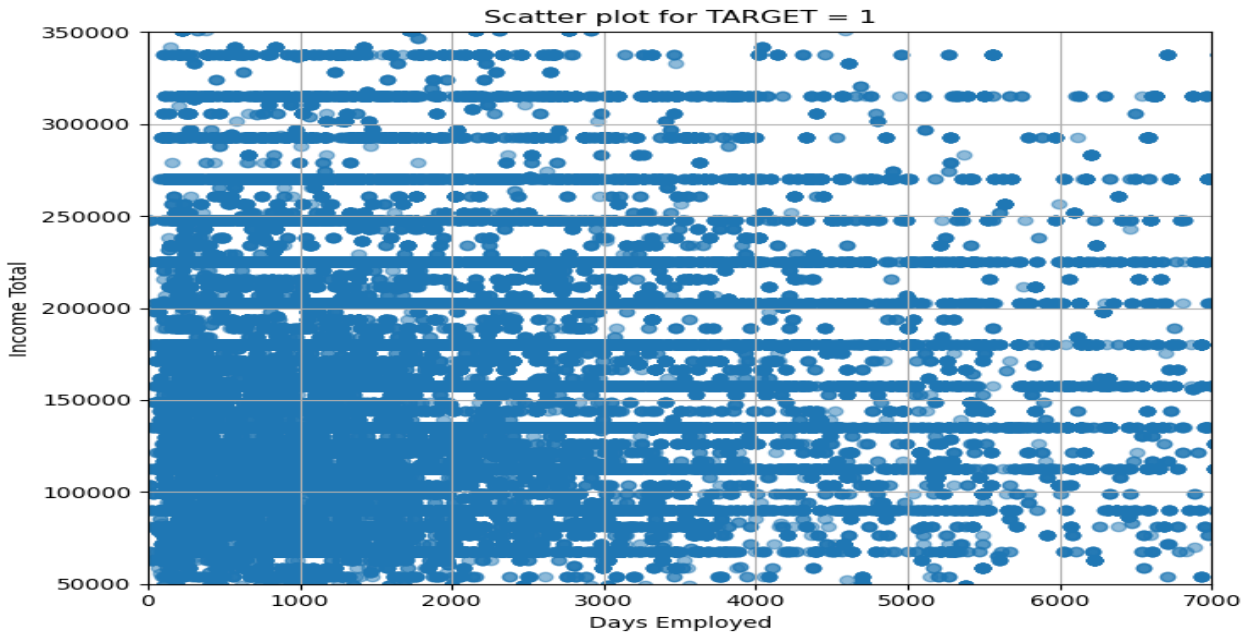


Multivariate Analysis

TARGET column comparison with DAYS_EMPLOYED & AMT_INCOME_TOTAL

Clients with higher DAYS_EMPLOYED had no payment difficulties, they also had higher income. Whereas clients' who had payment difficulties we see less population who were more than 3000 DAYS_EMPLOYED.

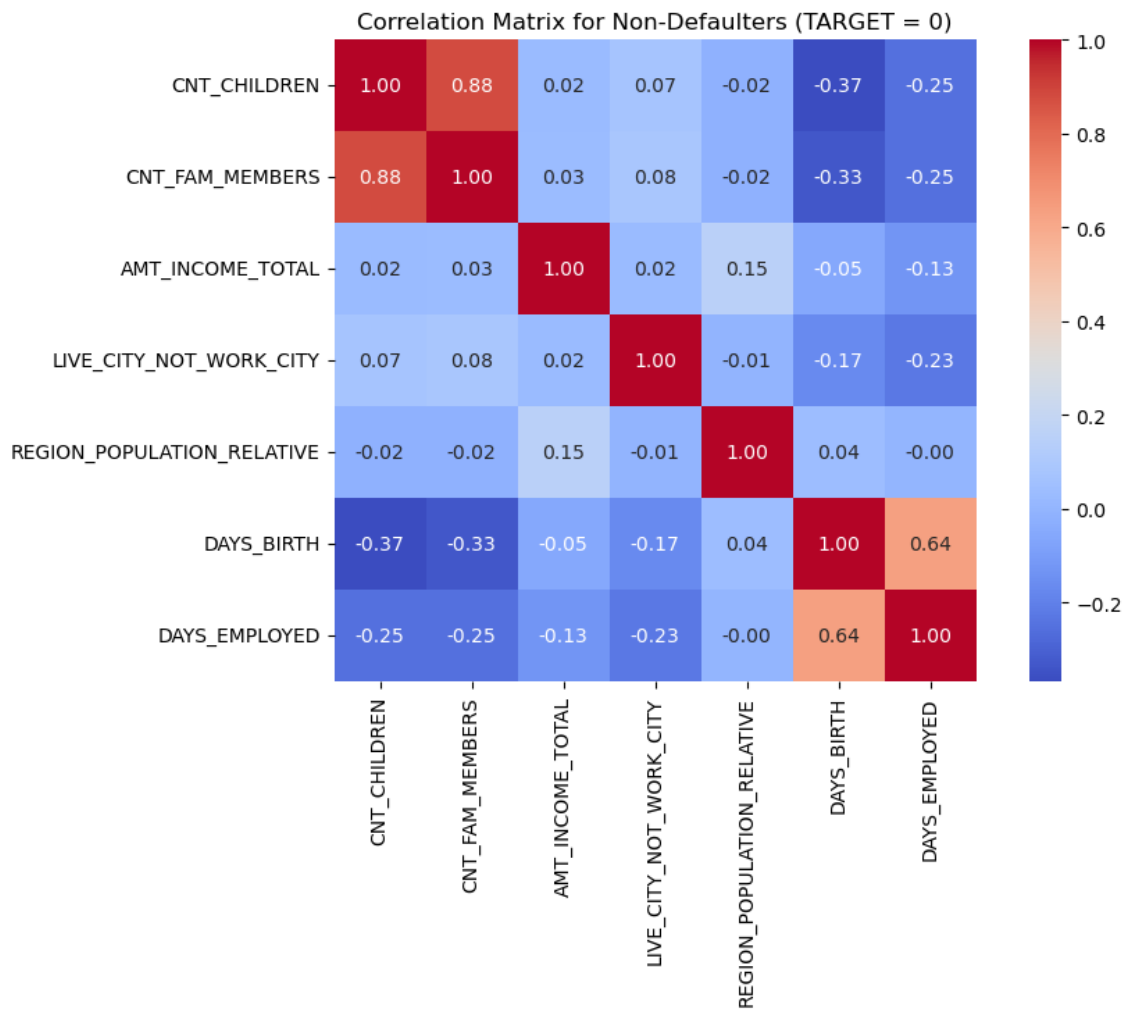




Multivariate Analysis

CORRELATION MATRIX

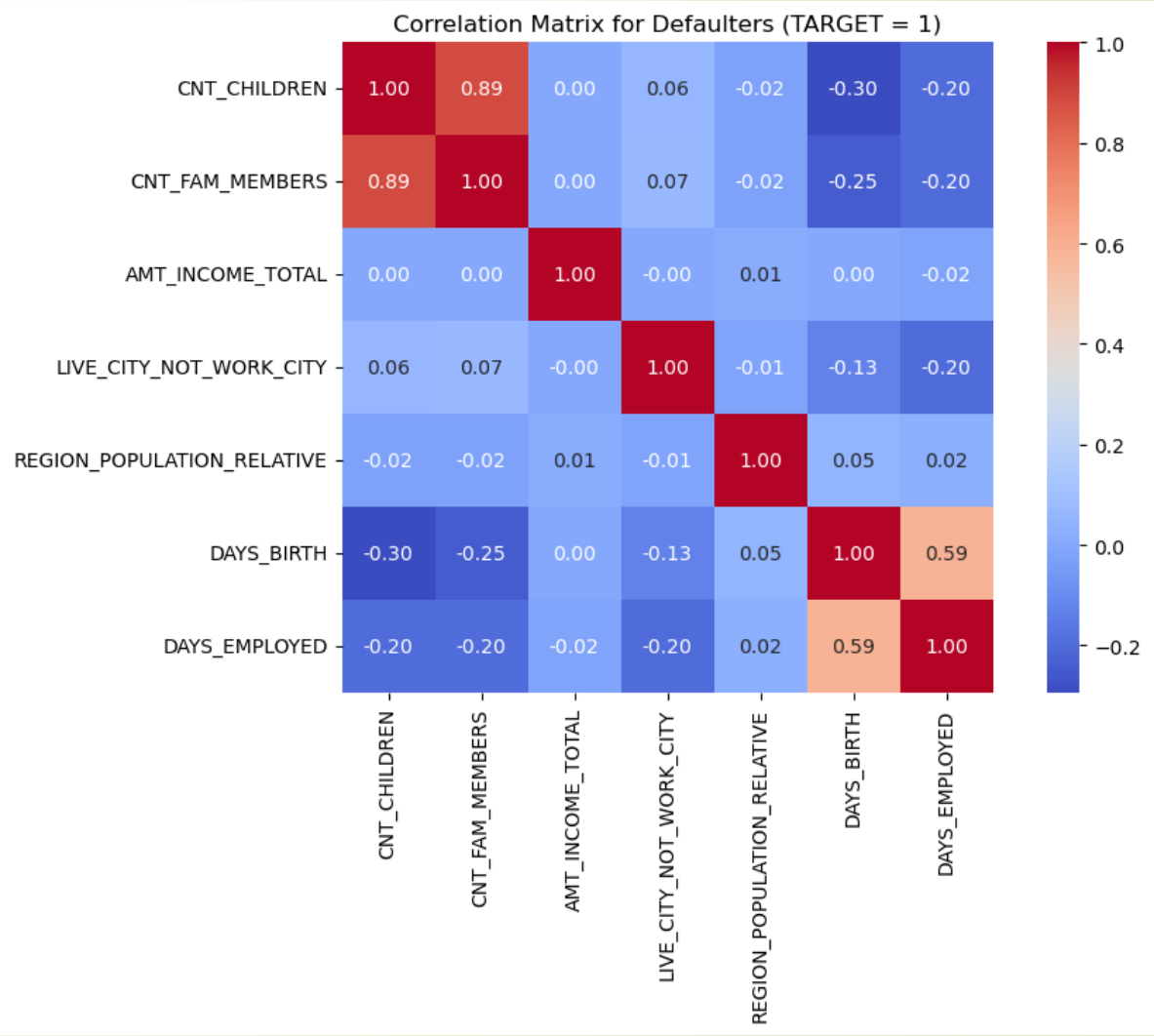
From the merged data we see strong correlation between **DAYS_BIRTH** and **DAYS_EMPLOYED** which is around **0.64** in **Non-Defaulters**



Multivariate Analysis

CORRELATION MATRIX

From the merged data we see strong correlation between **DAYS_BIRTH** and **DAYS_EMPLOYED** which is around **0.59** in **Defaulters**



INSIGHTS

NAME_CONTRACT_TYPE: Most of the clients choose Cash loans or Consumer loans .
CODE_GENDER : Female populations took more loan then male, though they earned lesser than males. Hence, they are likely to face more payment difficulties.
FLAG_OWN_CAR : most of the population did not own a car.
FLAG_OWN_REALTY: three fourth of the population did own a flat or house.
CNT_CHILDREN : population with no children were mostly seen taking loans.
AMT_INCOME_TOTAL : Majority of the population had income range in 76 to 113K .
AMT_CREDIT : Most of them took credit in the range of 222 to 259K.
AMT_ANNUITY: 80 to 90% of the population had annuity in the range of 40 to 76K.
AMT_GOODS_PRICE: Several of the population bought goods priced with in the range of 222 to 259K.
NAME_TYPE_SUITE: Many of them were "Unaccompanied" while taking loans.
NAME_INCOME_TYPE: More number of **"**working" group have taken the loan and are likely to default too.
OCCUPATION_TYPE : Laborers are found to default more .
CNT_FAM_MEMBERS: population who have around 2 members are more likely to default.

INSIGHTS

NAME_EDUCATION_TYPE: Secondary and secondary special have taken more loans and are more likely to default.
NAME_FAMILY_STATUS: More number of married population have taken loan and are likely to default too.
NAME_HOUSING_TYPE: population who have house/apartment have taken loans and more likely to default.
DAYS_EMPLOYED: More number of population have worked less than 3000 days or 8 years.
ORGANIZATION_TYPE: people who work in "Business Entity 3" have taken more loan and are more likely to default.
NAME_PAYMENT_TYPE: lot of them pay the bank by "Cash through the bank"
NAME_CLIENT_TYPE: Majority of Repeaters have taken loan.
AGE or DAYS_BIRTH : Majority of the population within the age group of 31-40 have taken loans and they are found to default more when compared to 41-50 and 51 – 60.

INSIGHTS

Clients who can pay back loan on time OR Bank to target these population for **approving** loans.

1. **Clients who have income range 76 to 186K**
2. **Clients who are in the age group of 21-30 and 41-50 .**
3. **Clients who have higher education and incomplete higher.**
4. **Clients who are employed in an organization for more than 4000days or 10 years .**
5. **Clients who come back after paying the initial loans , Repeaters.**
6. **Clients who have occupation such as Core staff and Sales staff .**
7. **Clients who work in organization such as Medicine, Others and are Self Employed.**
8. **Clients who are single or had a civil marriage .**
9. **Clients who do not own a house or flat.**
10. **Male population with more income range.**

INSIGHTS

Clients who can not pay back loan on time OR Bank to target these population for **refusing loans.**

1. **Clients who have more than 2 number of children or family members.**
2. **Academic degree holders are more likely to default.**
3. **Clients who are managers and laborers.**
4. **Clients who work in Business Entity 3 organization type .**
5. **Clients who are married or separated.**
6. **Female population with less income.**

Thank you