CREDIT EDA CASE STUDY

By:

Saptarshi Ganguly

Nithyashree PV

Handling Missing Values

CATEGORICAL VARIABLES

Mode (most popular category)

Ex: Name Type Suite

Set as Null Values

(for smaller percentage of missing values)

Ex: Code Gender

Treat as Missing Values

(for significant percentage of missing values)

Ex: Occupation Type

NUMERICAL VARIABLES

Median (for values with many outliers)
Ex: Amount Annuity

Mode (for better measure of central tendency):

Ex:Amount Goods Price,EXT_SOURCE_3

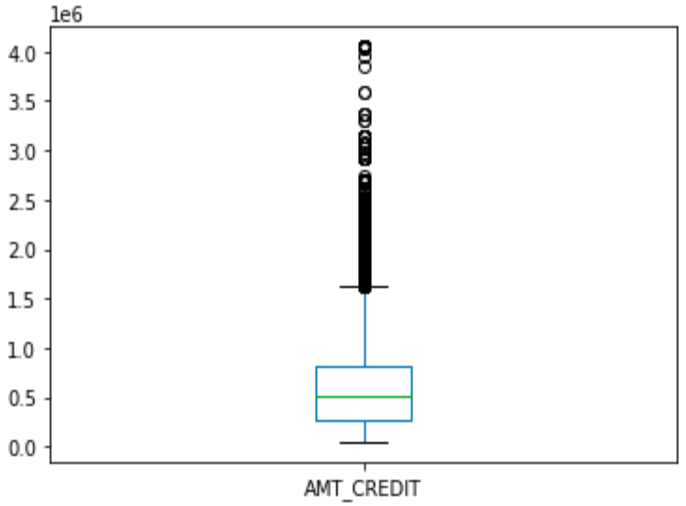
Mean(for no outliers, and smaller percentage of missing values)

Ex: EXT_SOURCE_2

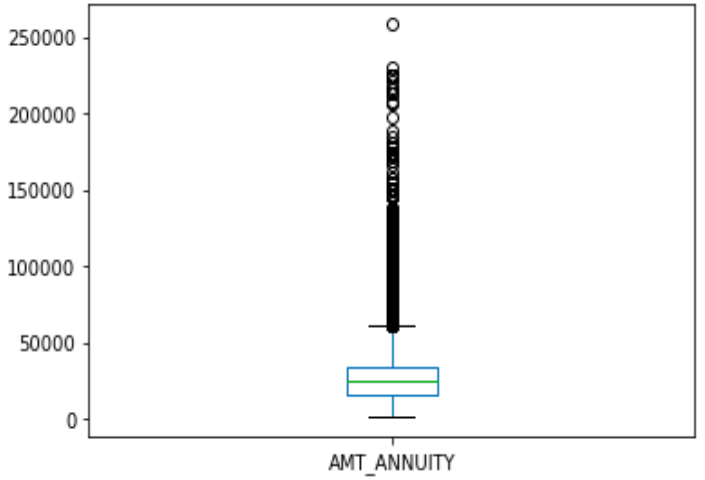
We have dropped unnecessary columns and those with over 50% missing values.

We have recommended dropping records with invalid entries. Ex: Amount Income Total

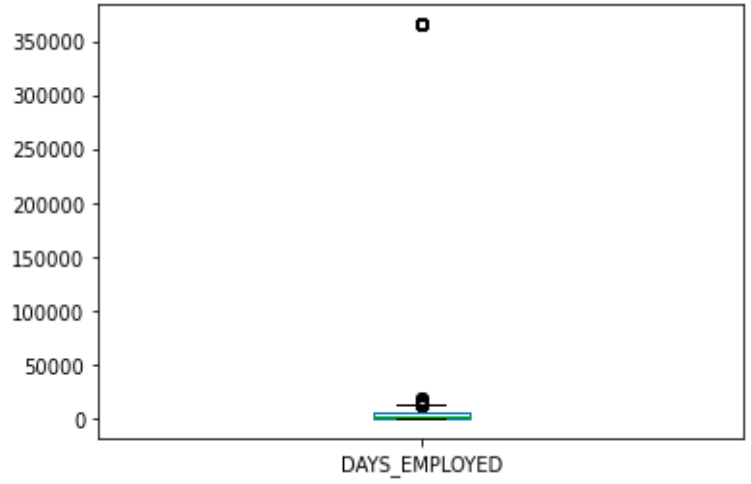
Handling Outliers



Capping at 99 percentile for outliers that are valid entries

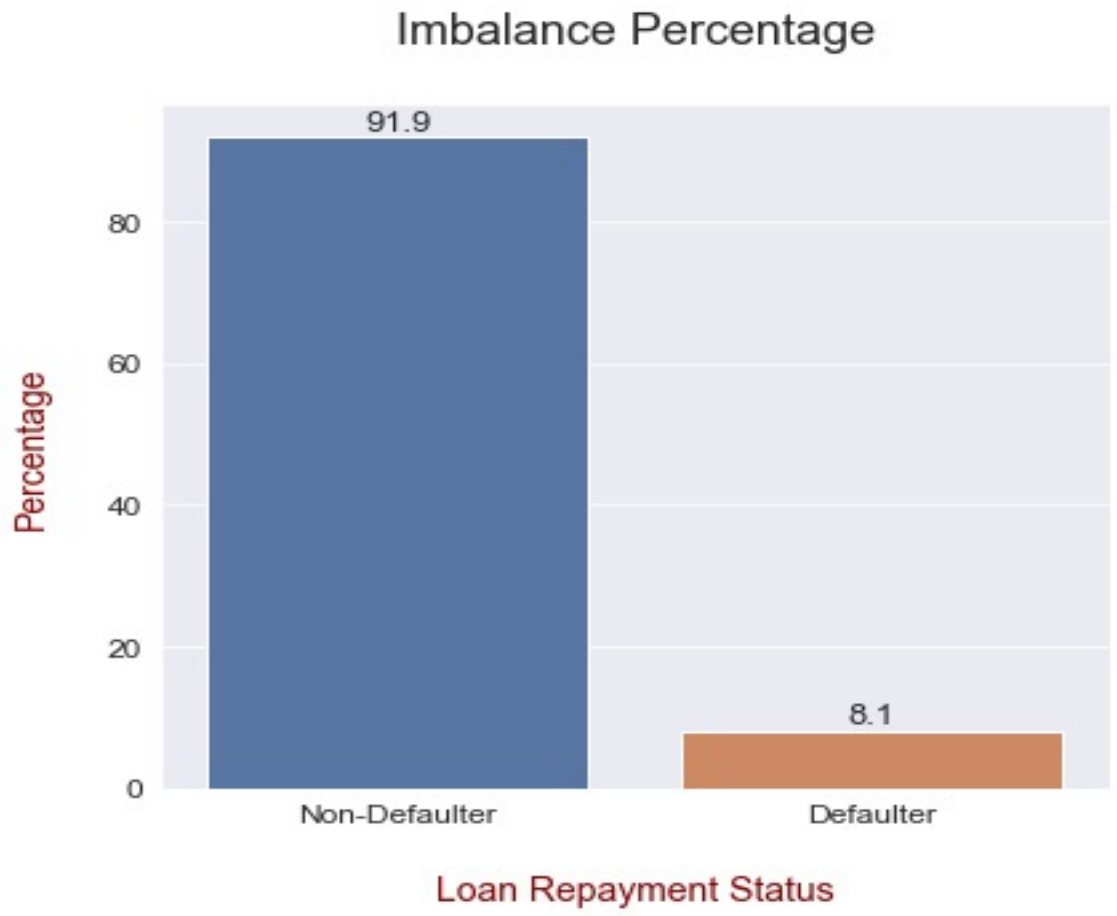


Binning values due to the many outliers present. And capping for the extreme outliers



Converting to null values for significant number of records with incorrect/invalid entries

Imbalance Analysis



Inference:

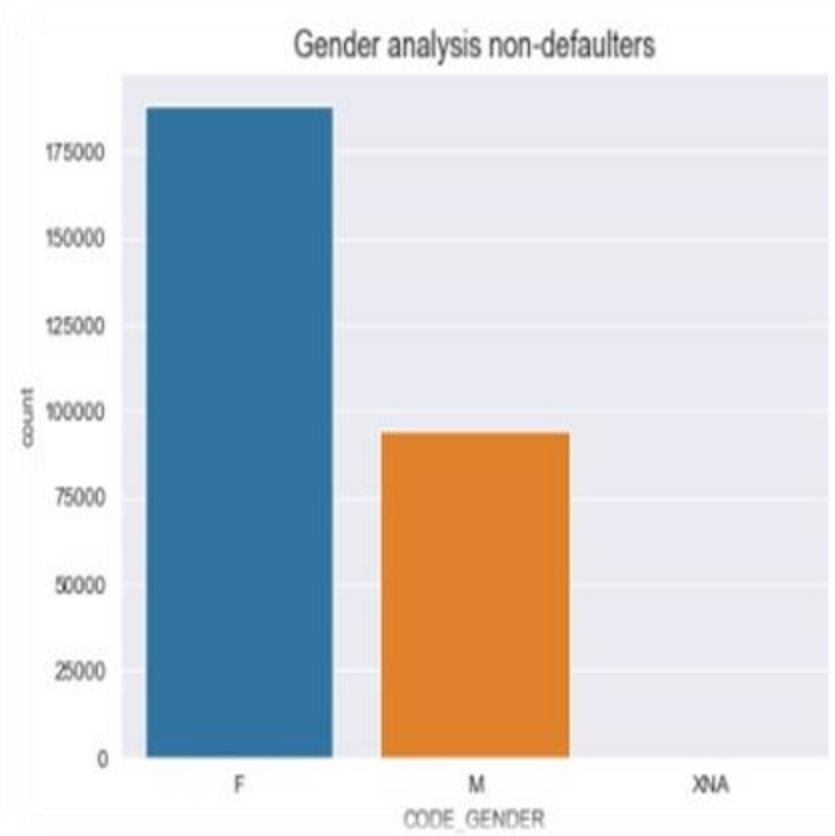
-There is a stark imbalance for the Target variable.

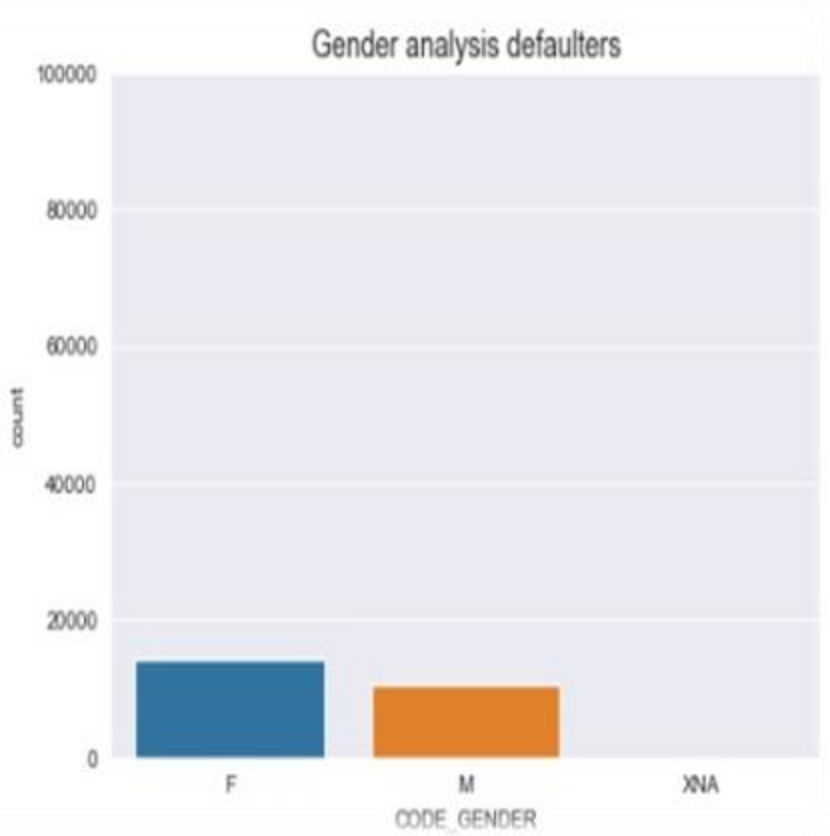
- The percentage of Non-Defaulters is almost 92% and that of Defaulters is 8% for the application data.

- The Ratio of Imbalance for Target Variable is 1 : 11.39.

Categorical Univariate Analysis

GENDER





Inference:

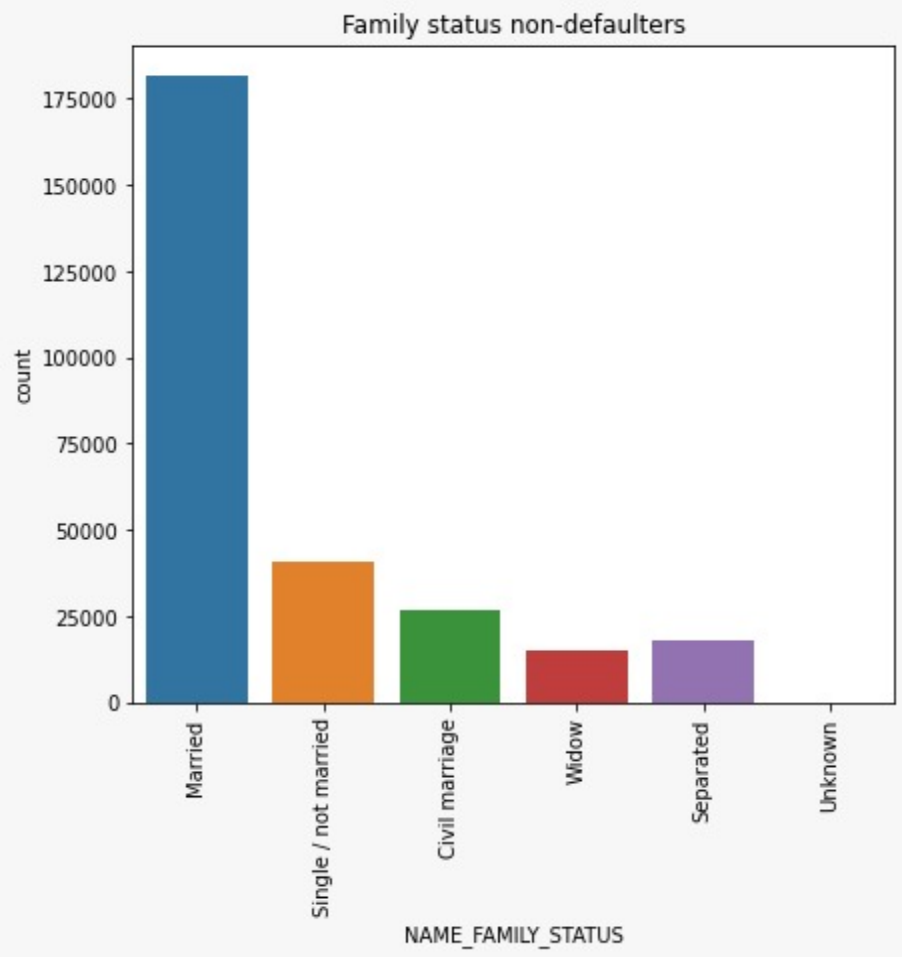
-65% of applicants are female.

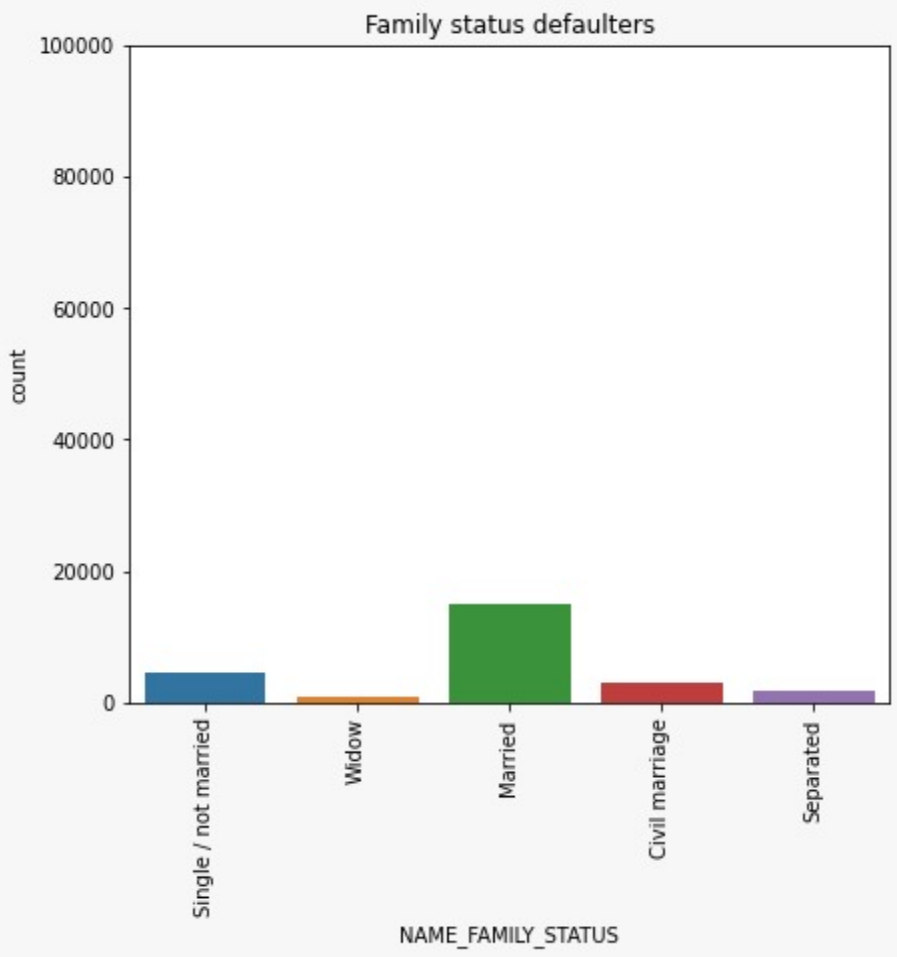
-The percentage of defaulters is higher among male male applicants, than among female applicants.

- Women are less likely to be defaulters.

Categorical Univariate Analysis

FAMILY STATUS





Inference:

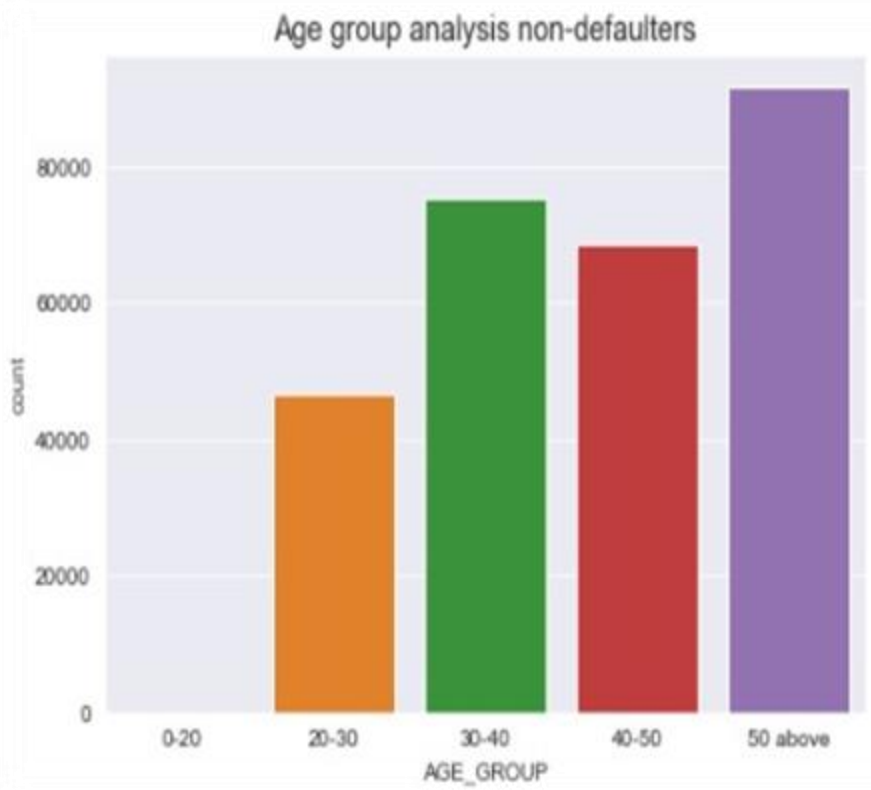
- Married, Single/Not married & Civil Marriage are the most common status.

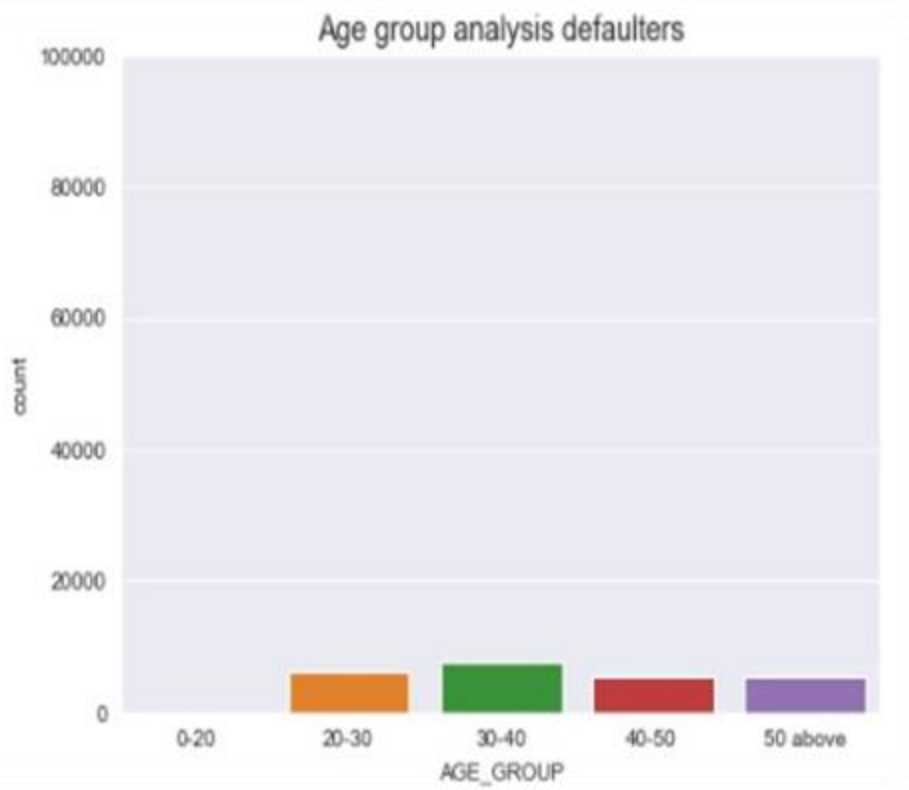
- Civil marriage and Single/Not married have the highest percentage of defaulters.

- Widows and Married have lowest percentage of defaulters.

Categorical Univariate Analysis

AGE GROUP





Inference:

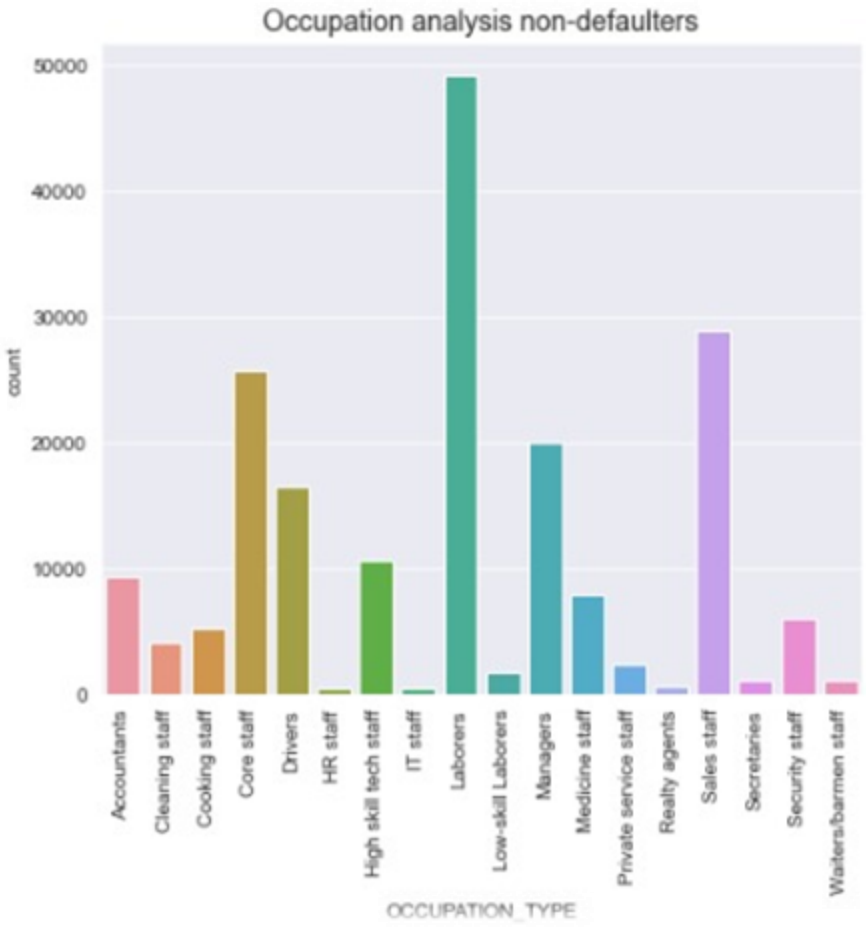
-30-40 and 20-30 age group have the highest percentage of defaulters.

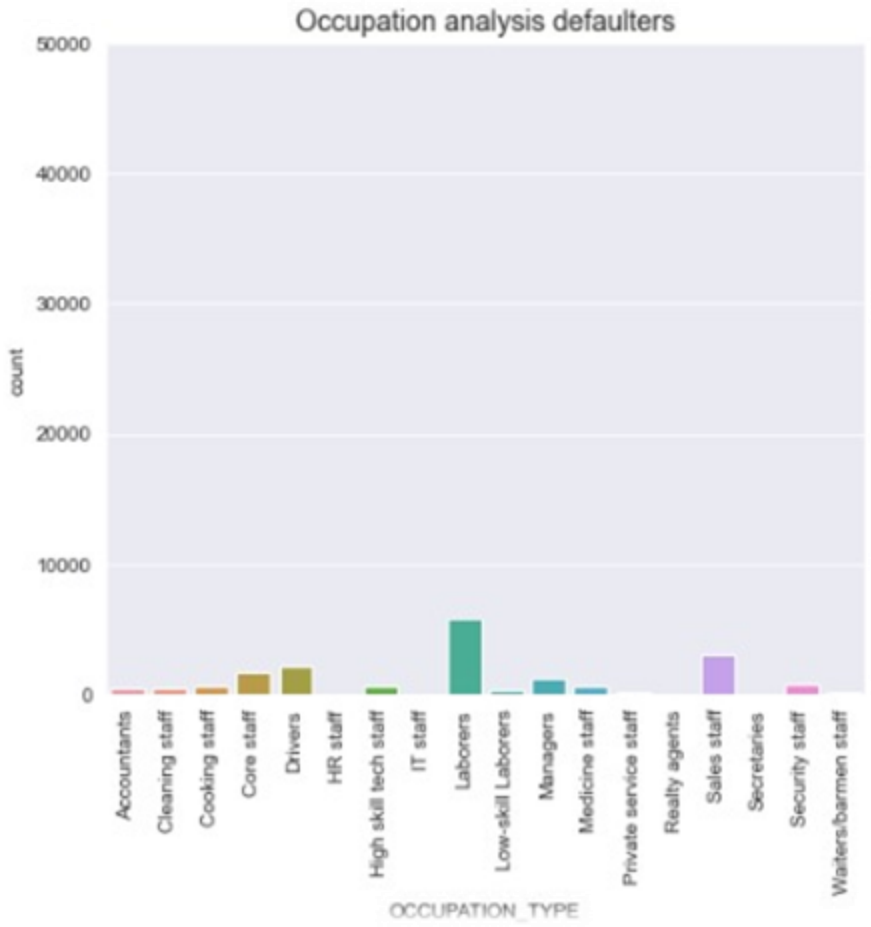
- 50 and above are the age group with largest percentage of non defaulters.

- 50 and above are 31% of total loan applicants. 30-40 age group at 27%.

Categorical Univariate Analysis

OCCUPATION TYPE





Inference:

-Most applicants are Laborers. Followed by Sales staff and Core staff.

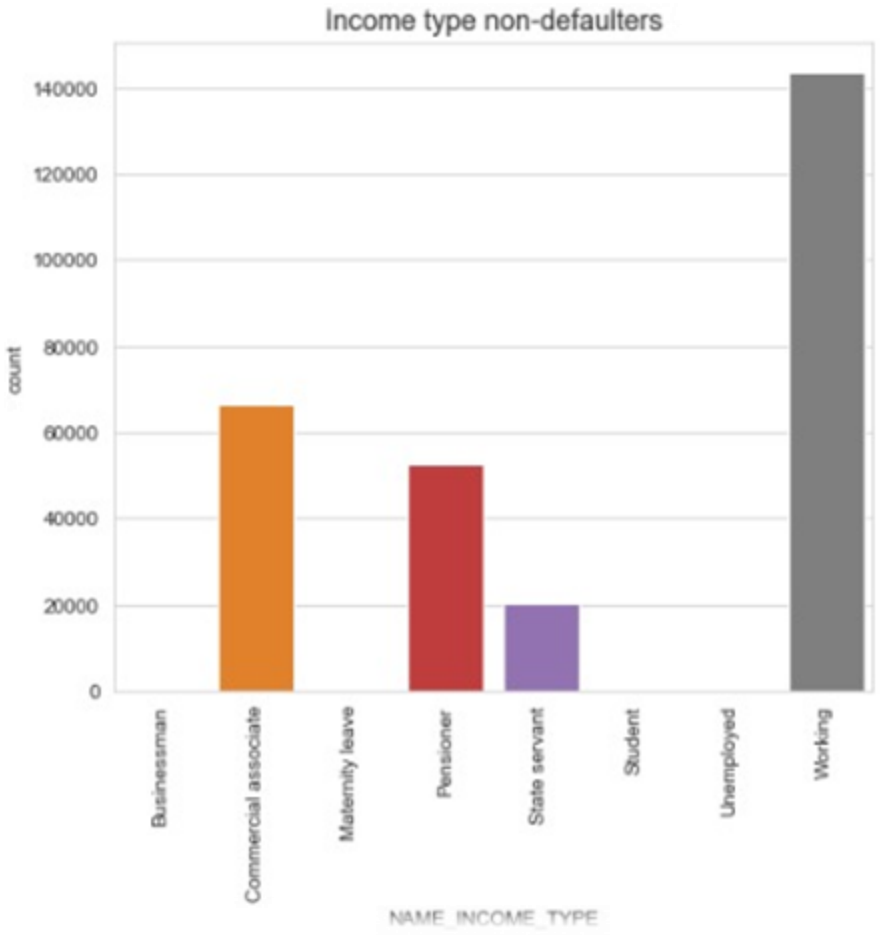
- Low Skill Laborers, Drivers, and Waiters have high percentage of defaulters.

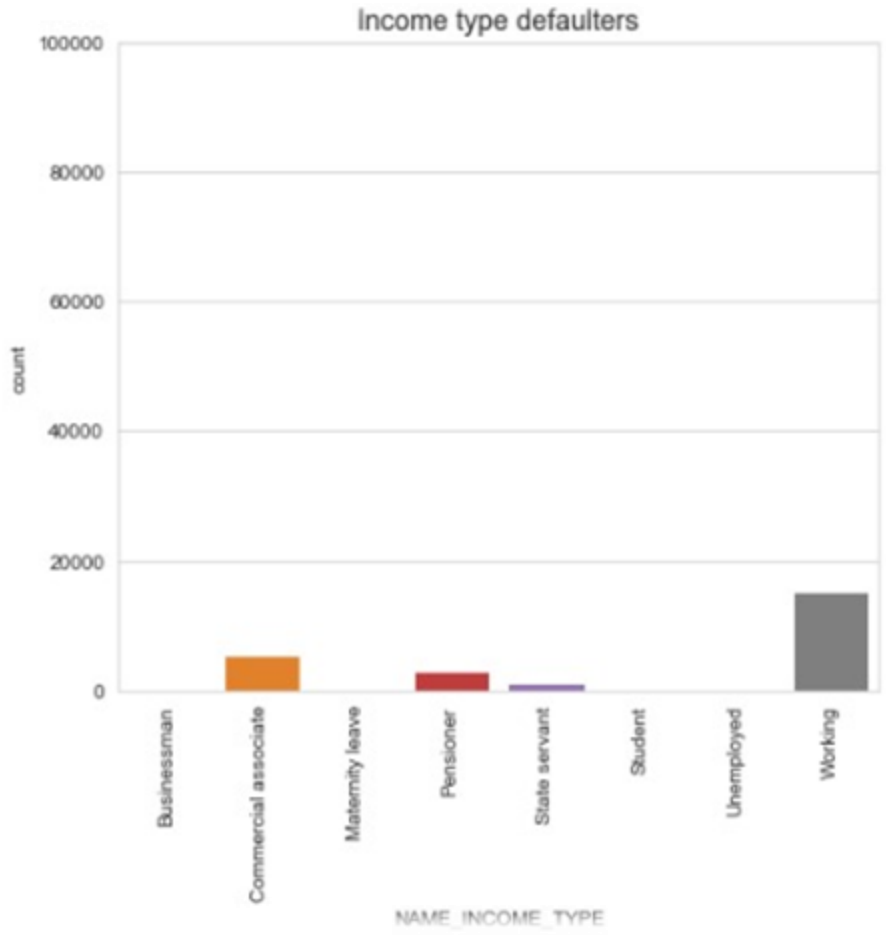
- Student and Businessman are few and have no defaulters.

- Accountants, High Skill tech staff and Managers have low percentage of defaulters.

Categorical Univariate Analysis

INCOME TYPE





Inference:

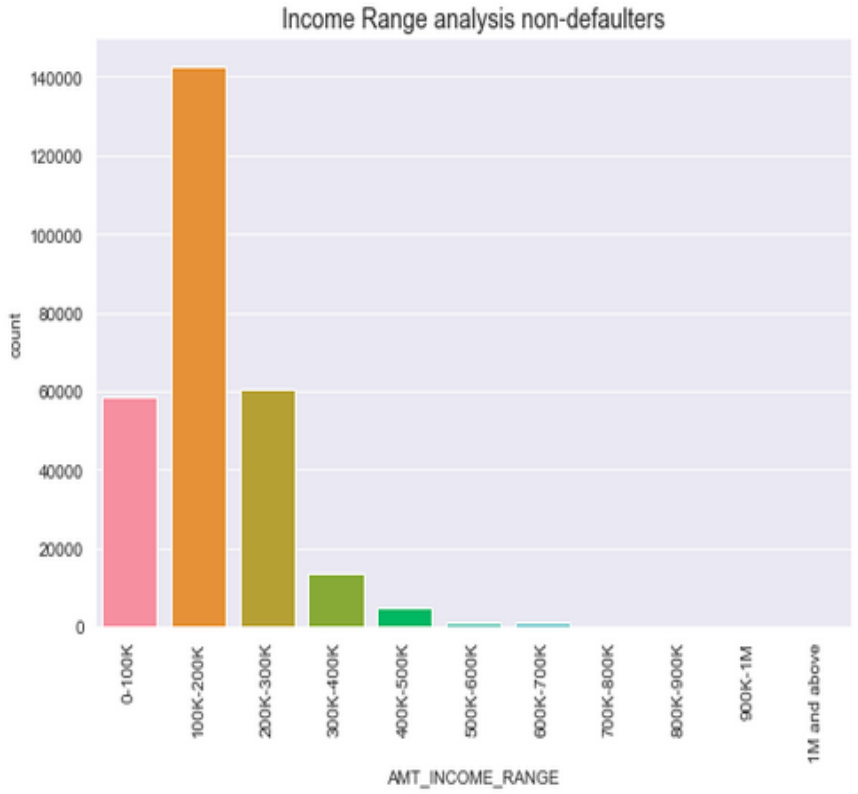
- The most common income types are:
- Working
- Commerical associate
- Pensioner
- State servants

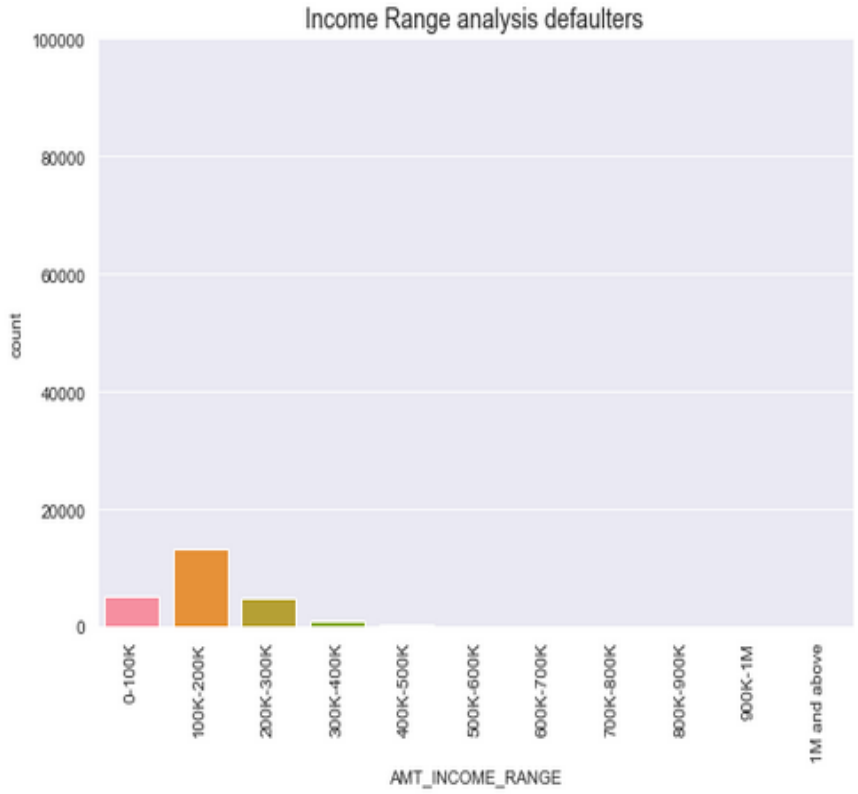
- Pensioners are the least likely to default. Followed by State Servants.

- Working, commerical associates have higher percentage of defaulters.

Categorical Univariate Analysis

INCOME RANGE





Inference:

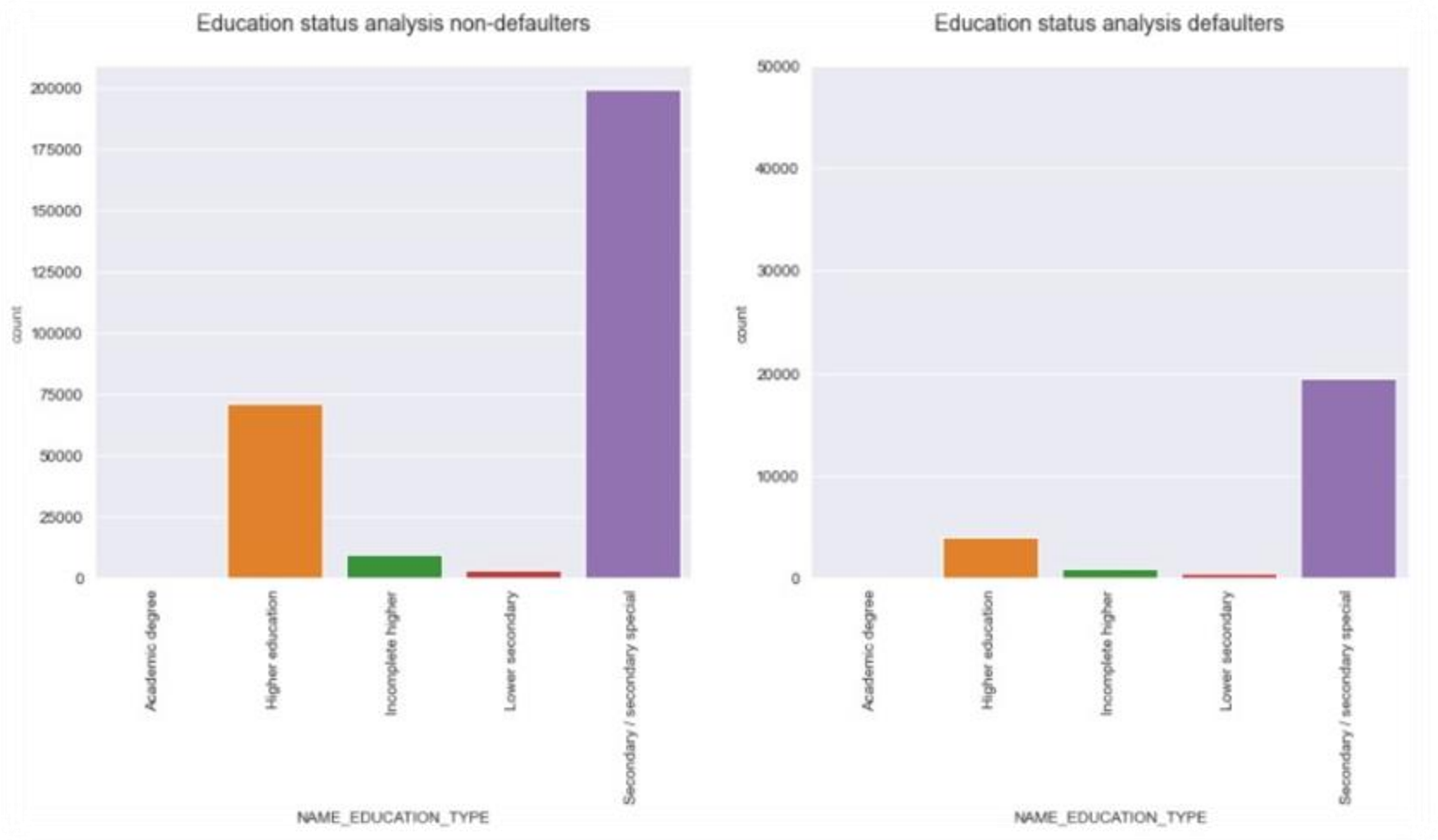
- Applicants earning between 100-200K are most likely to default.

- Applicants earning between 900K-1 million are the fourth most common defaulters.

- 90% of the applicants have an income less than 300K.

Categorical Univariate Analysis

EDUCATION TYPE



Inference:

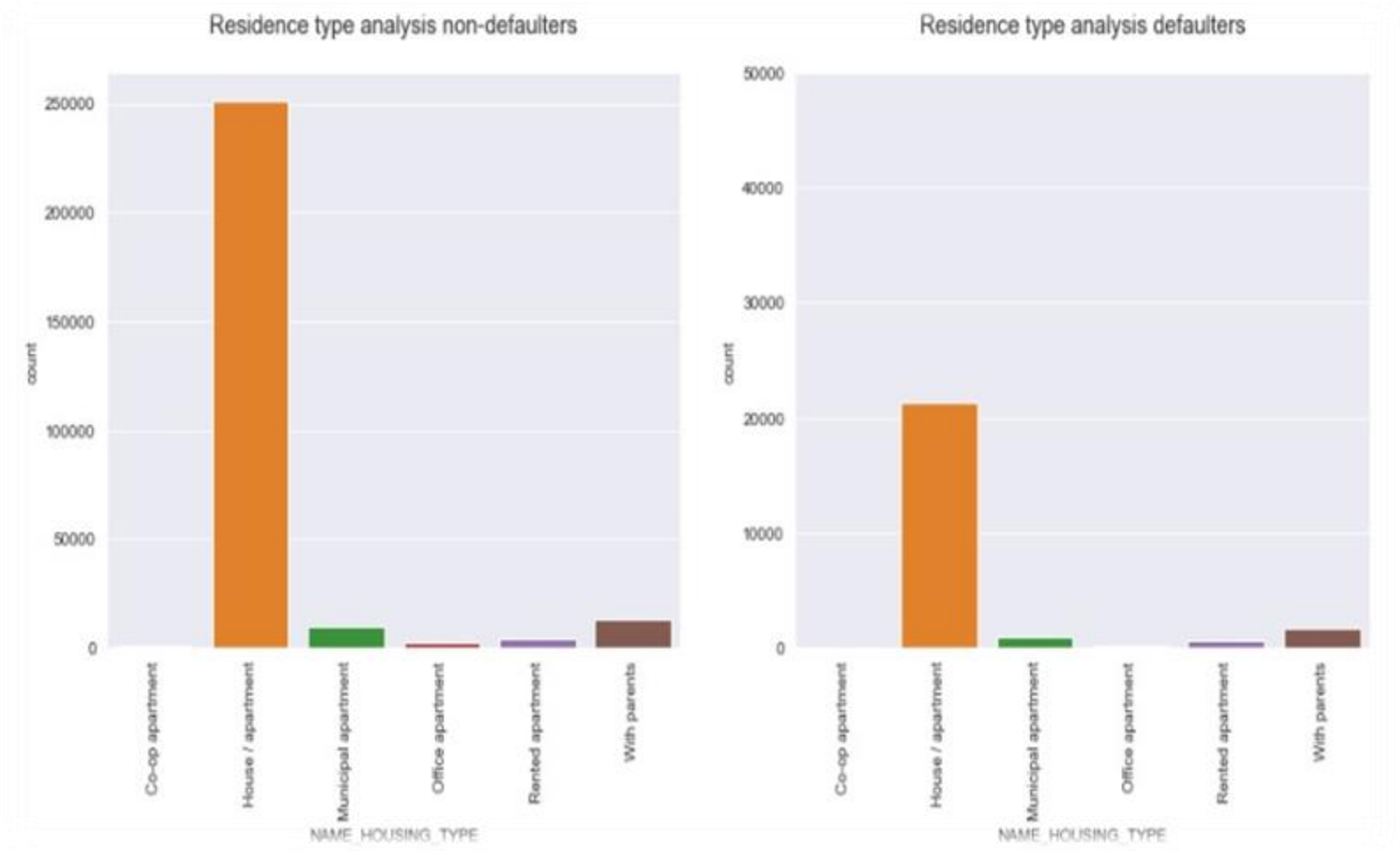
- Secondary education and Higher education are the most common.

-Higher education and Academic degree types have lowest default percentage

- Secondary education and Lower Secondary have the highest default percentage.

Categorical Univariate Analysis

RESIDENCE TYPE



Inference:

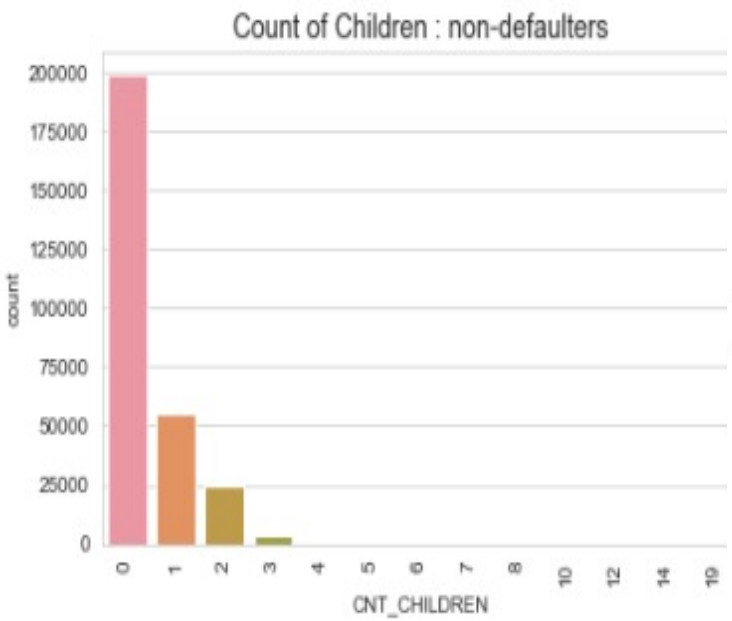
- House/Apartment and With parents are the most common type.

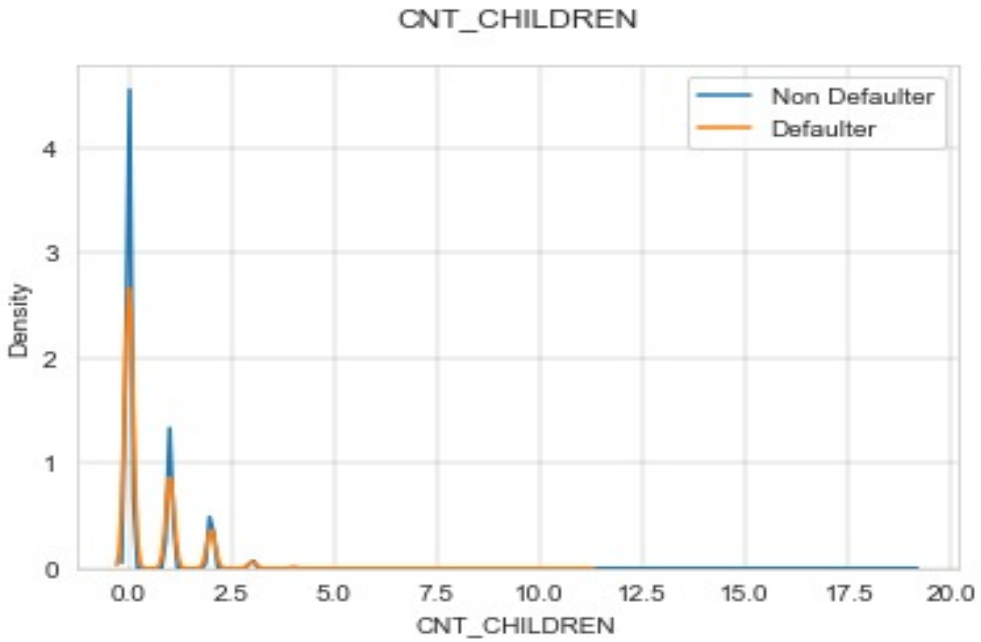
-With parents and rented apartment have the highest default percentage.

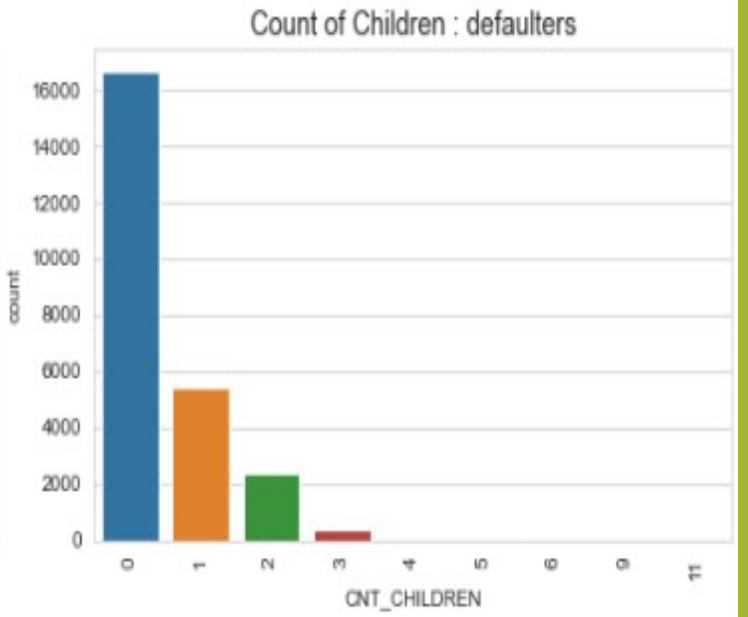
- House/Apartment, Office and Co-op apartments have the least defaulter percentage.

Numerical Univariate Analysis

COUNT CHILDREN







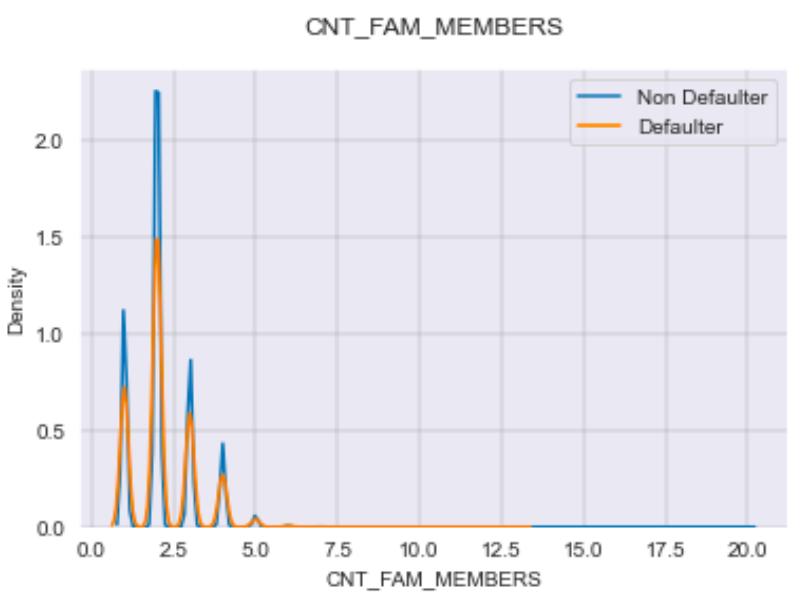
Inference:

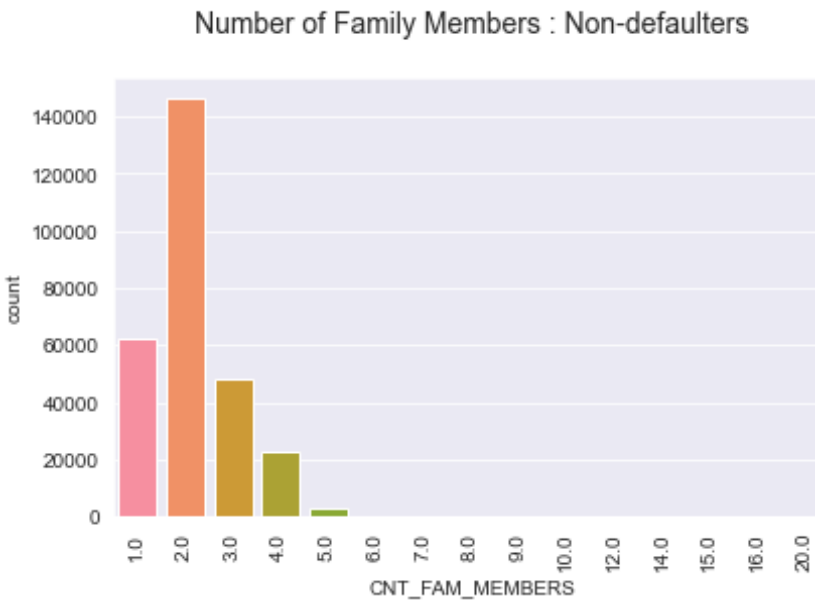
- 70% of applicants have no children. 98% have less than or equal to 2 children.

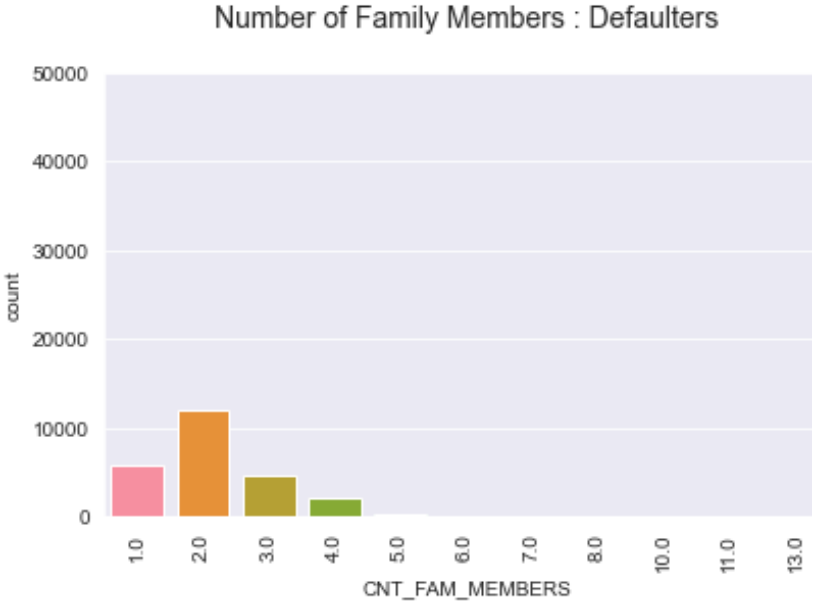
-Applicants with 3 or more children are more likely to default.

Numerical Univariate Analysis

COUNT FAMILY MEMBERS







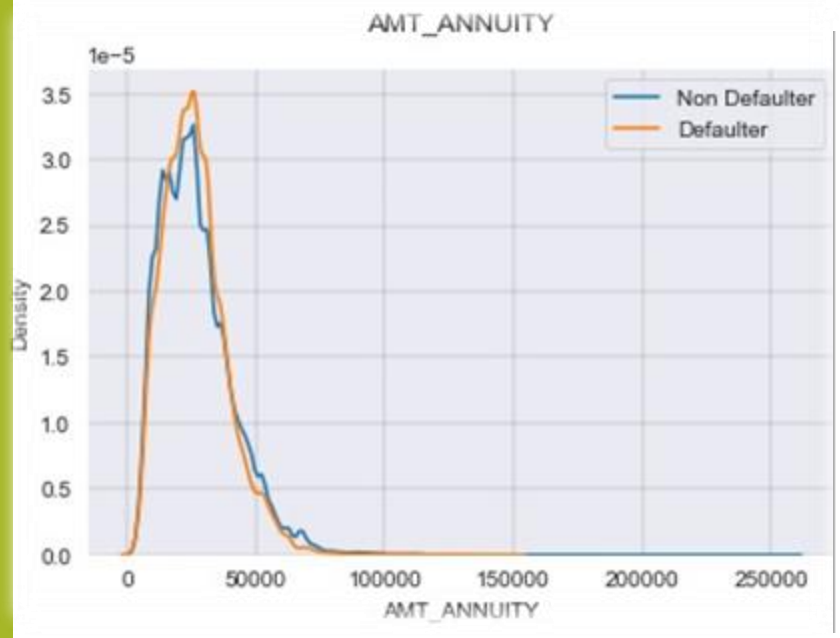
Inference:

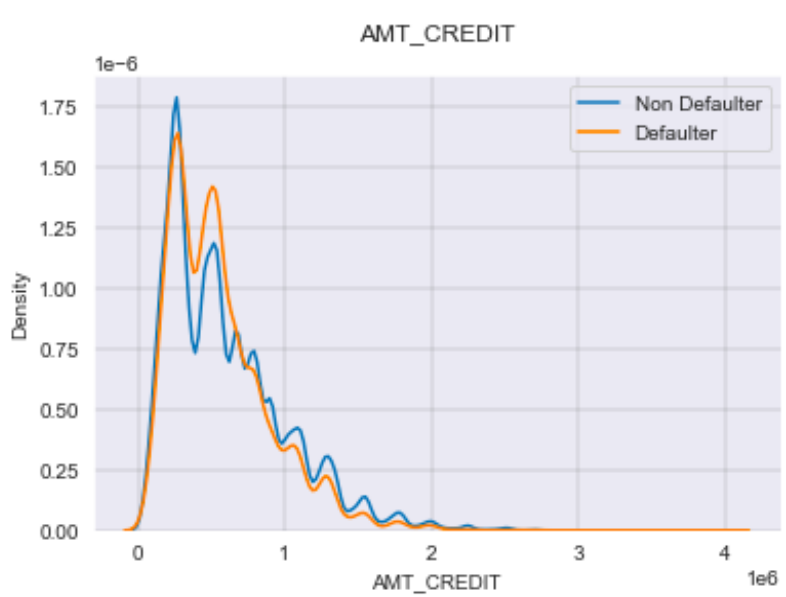
- 50% of applicants have 2 family members. 98% have less than or equal to 4 family members.

-Applicants with 3 –11 family members are more likely to default.

Numerical Univariate Analysis

AMOUNT (ANNUITY,CREDIT,GOODS PRICE)



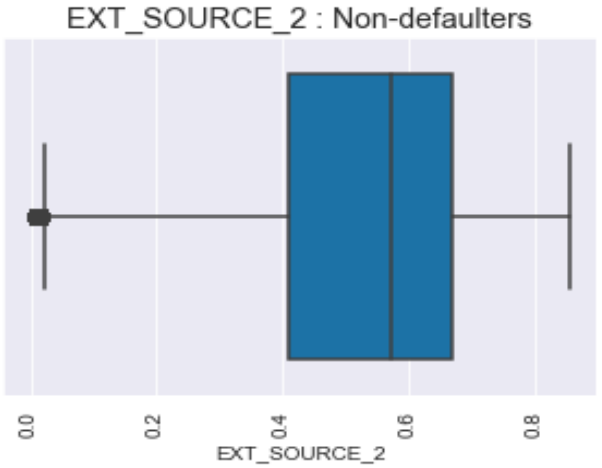


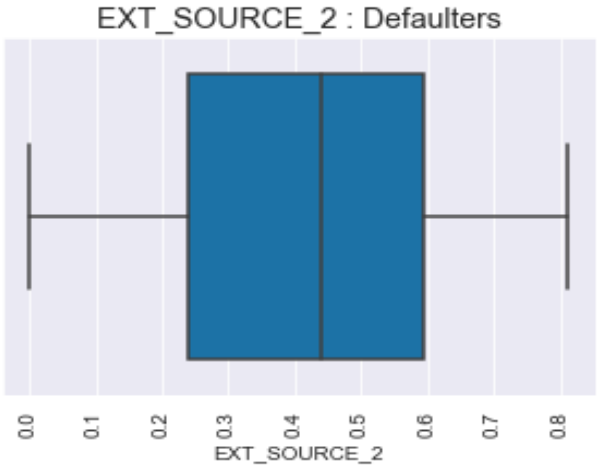
Inference:

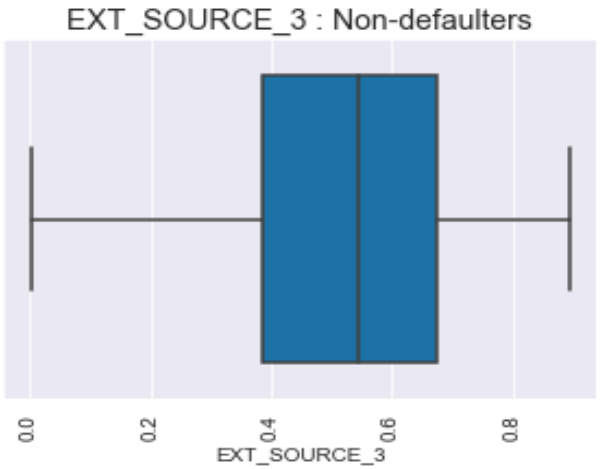
- The Amount Annuity paid by most applicants is less than 75,000, with majority below 50,000.Most defaulters have AMT_ANNUITY below 150,000.There are fewer defaulters with AMT_ANNUITY > 150,000.
- 80% of applicants have a credit loan less than 1 million. 200-300K credit range is the most common , and is second to 500-600K range in highest percentage of defaulters.
- The majority of the distribution of Goods Price among applicants is less than 1,000,000. Applicants with 400-600K and over 2.5 million amount goods price are have higher default percentage.

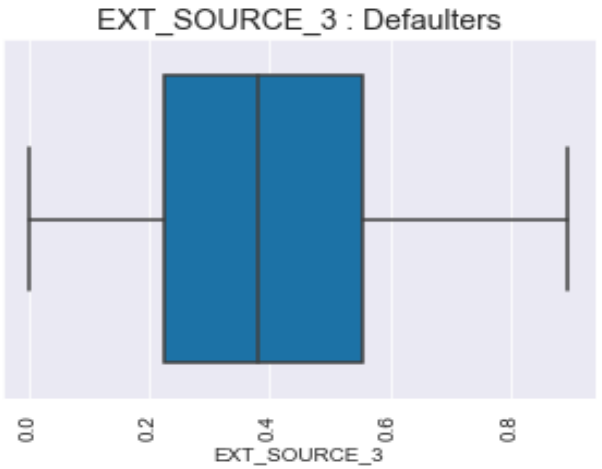
Numerical Univariate Analysis

EXTERNAL SOURCE SCORE









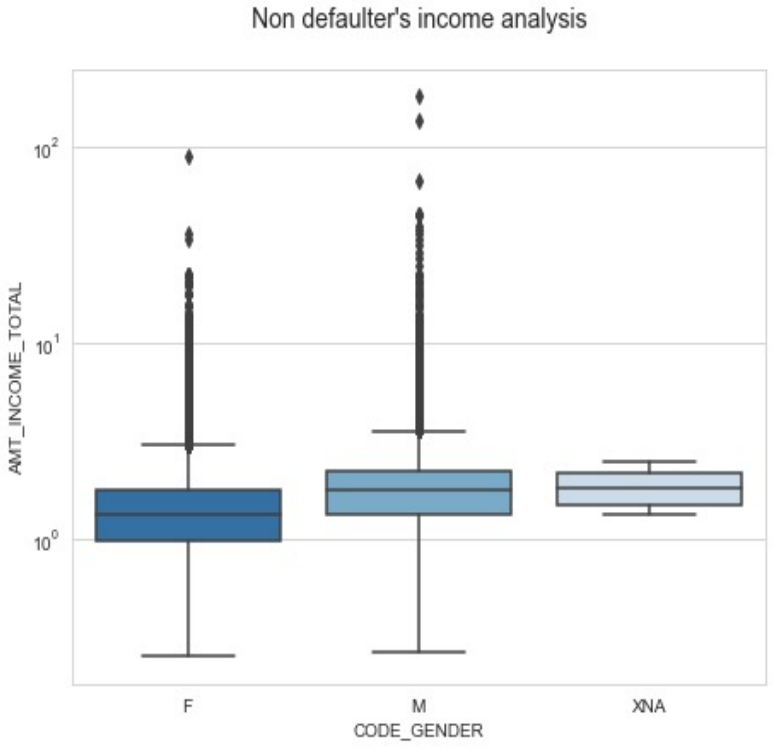
Inference:

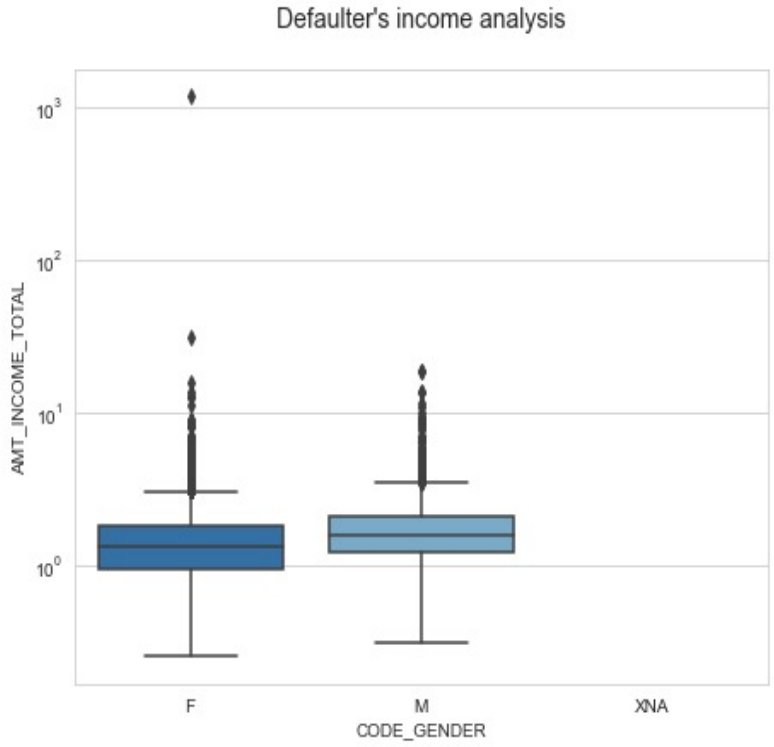
- Defaulters have a lower median score of 0.4 and a wider distribution ranging from a low 0.2 score.

- Non-defaulters have a median score greater than of equal to 0.5.

Bivariate Analysis

GENDER vs INCOME



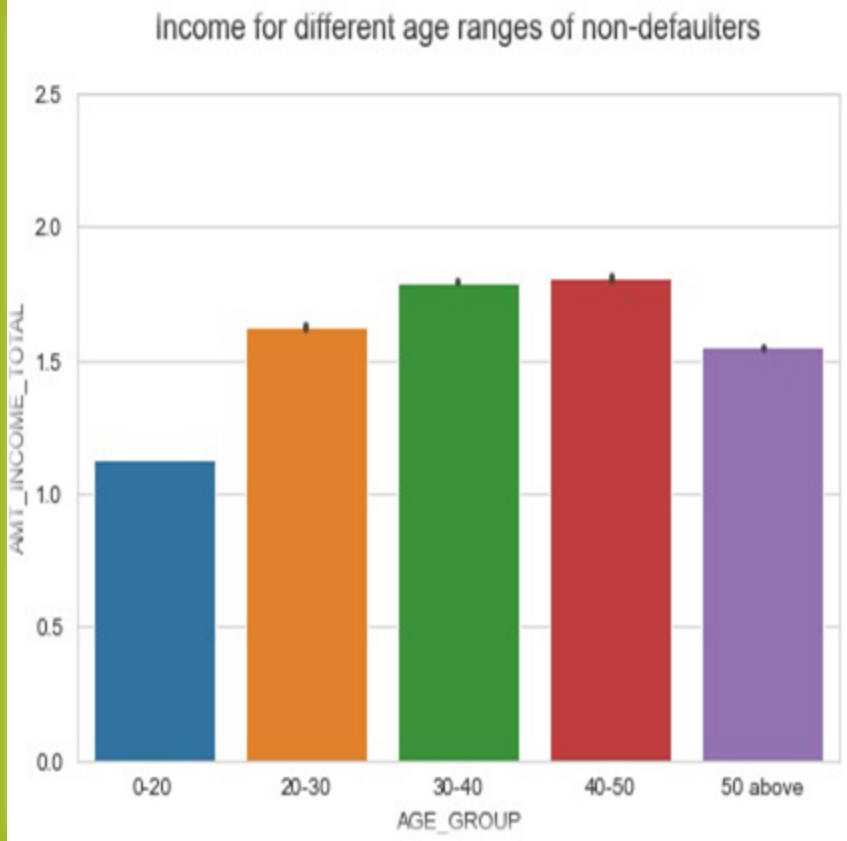


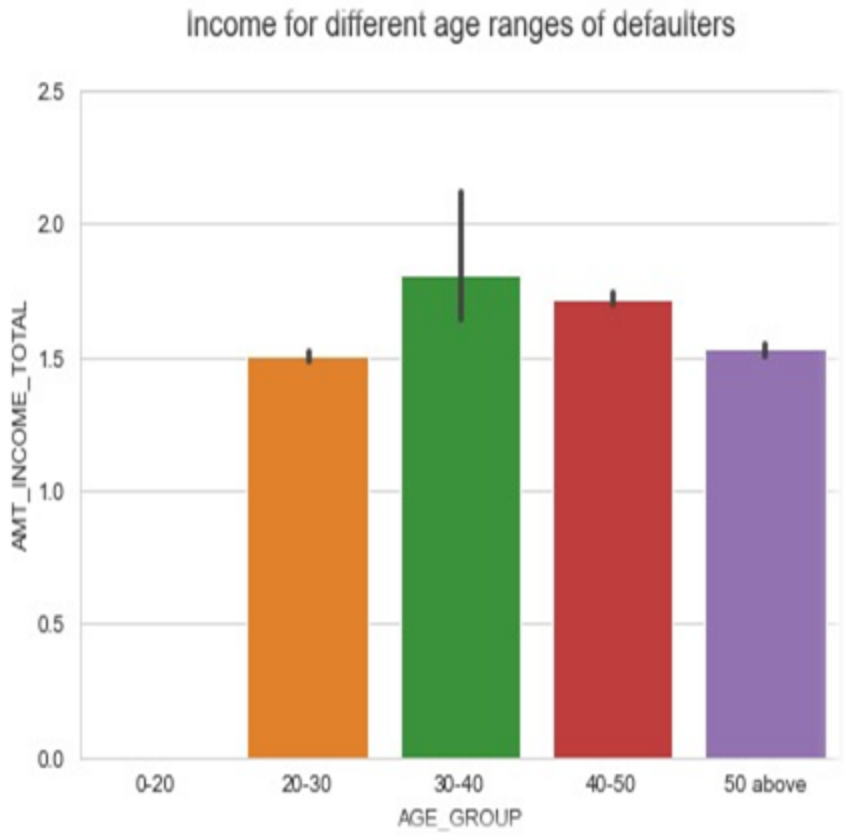
Inference:

-Male applicants are a smaller portion of the total, and have higher income compared to female applicants.
- Male applicants also have a wider range of income compared to female clients with more outliers present.

Bivariate Analysis

AGE vs INCOME





Inference:

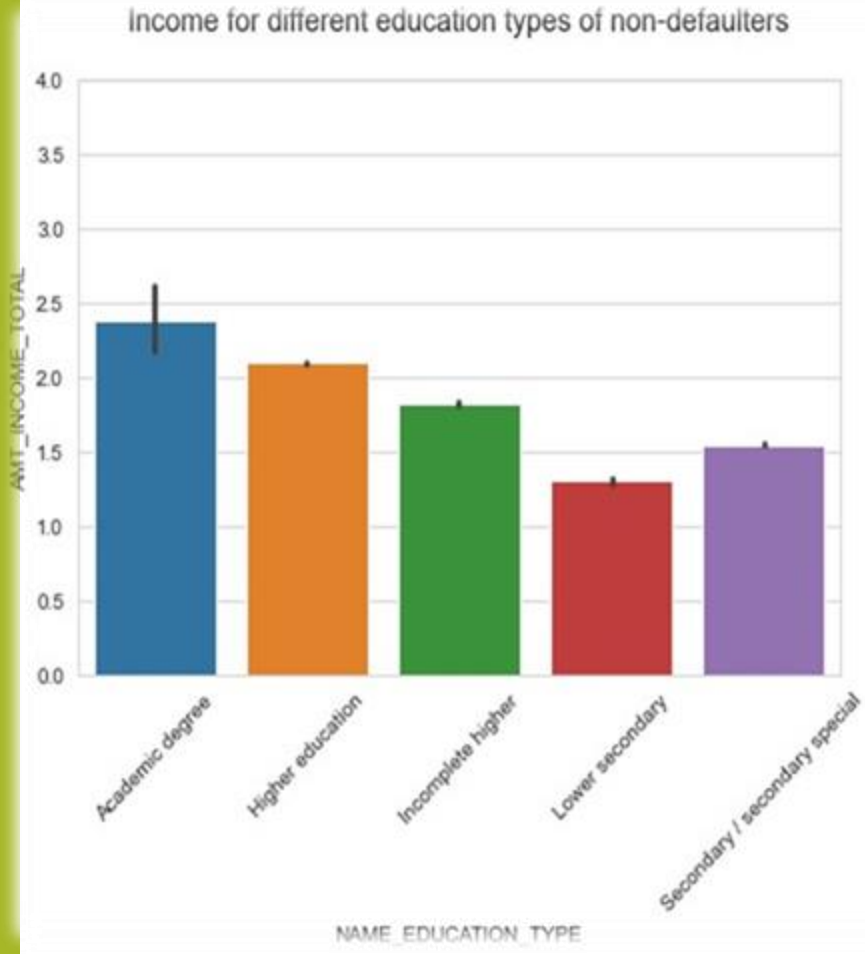
-The 30-40 age group has the highest amount income among the defaulters.

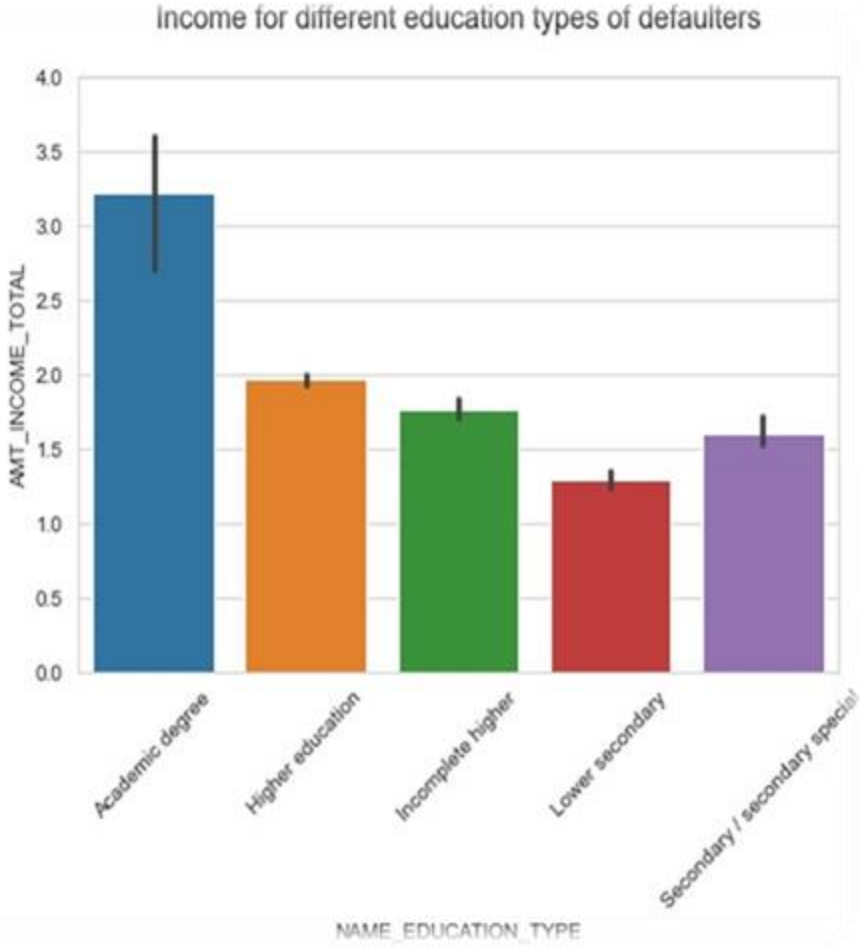
- Those with 40-50 have the highest income among non-defaulters, followed by 30-40.

- 30-40 group is 27% of total applicants, 40-50 is 24% of total applicants.

Bivariate Analysis

EDUCATION vs INCOME





Inference:

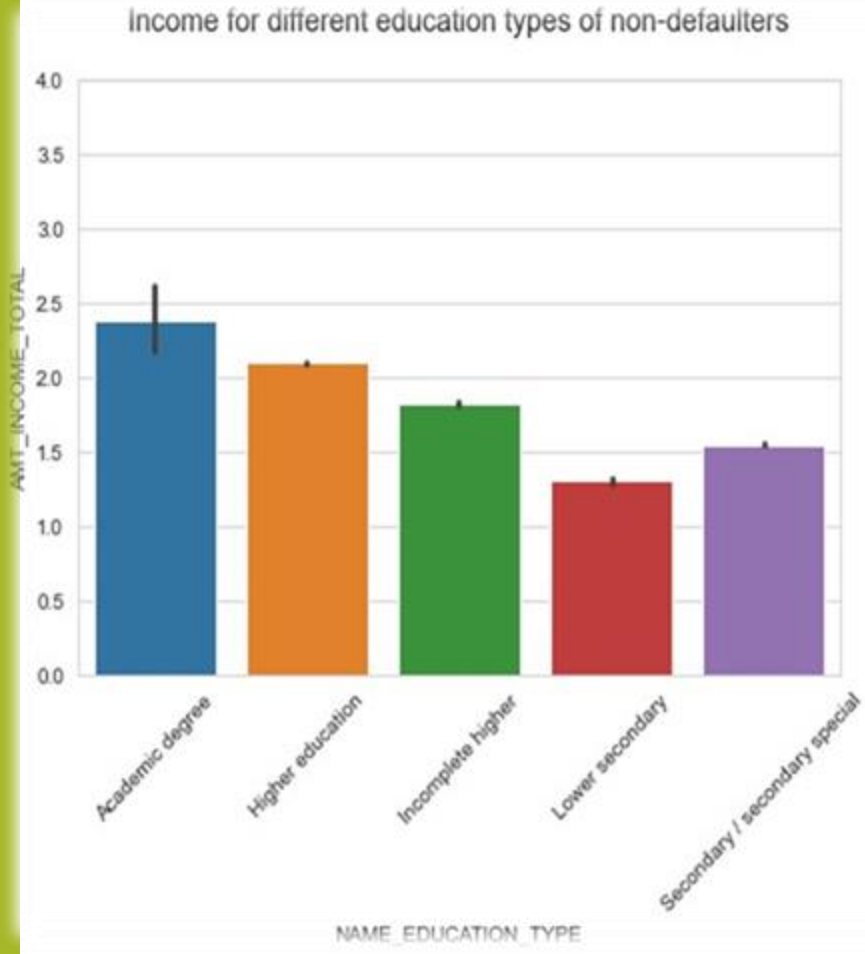
-Applicants with academic degrees have the highest income.

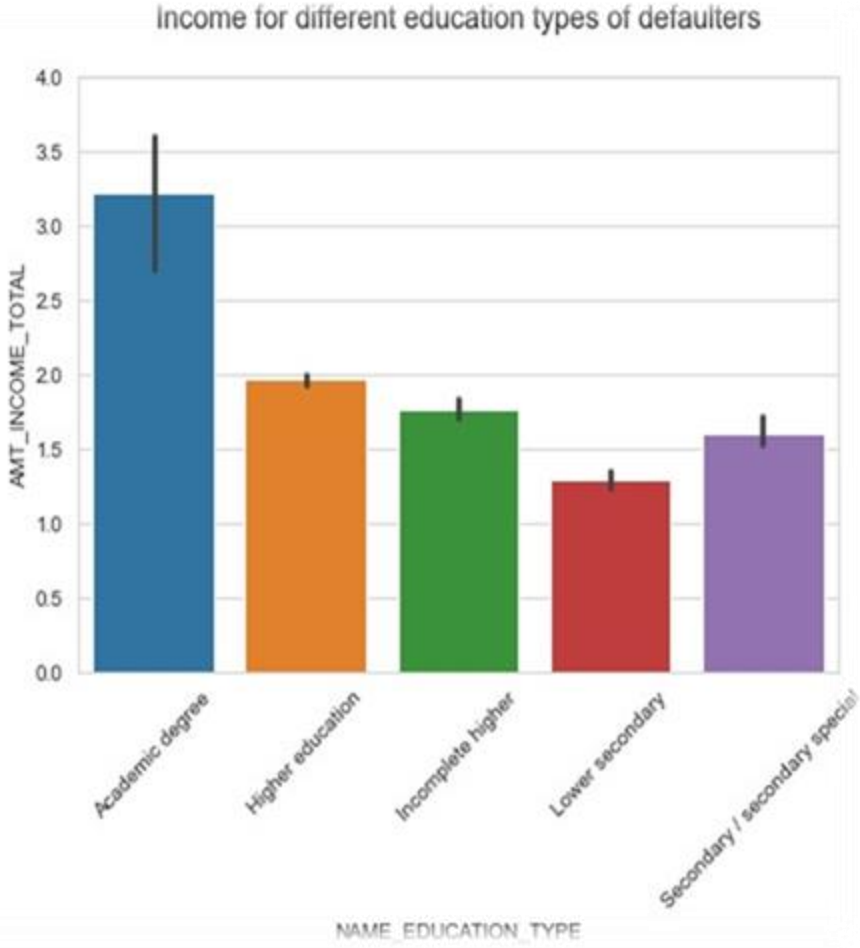
- Higher education is the second most common education and has the second highest income range among all applicants.

- Lower secondary and Secondary education have the lowest income.

Bivariate Analysis

EDUCATION vs INCOME





Inference:

-Applicants with academic degrees have the highest income.

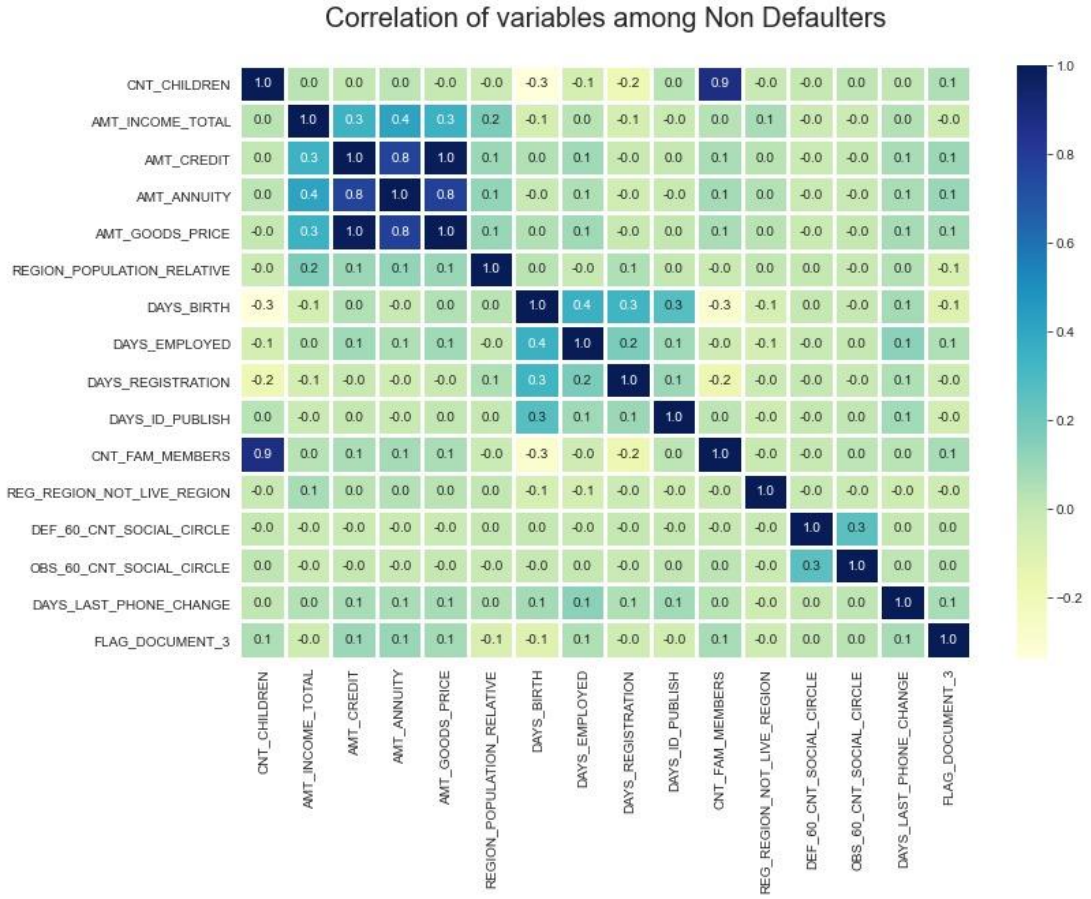
- Higher education is the second most common education and has the second highest income range among all applicants.

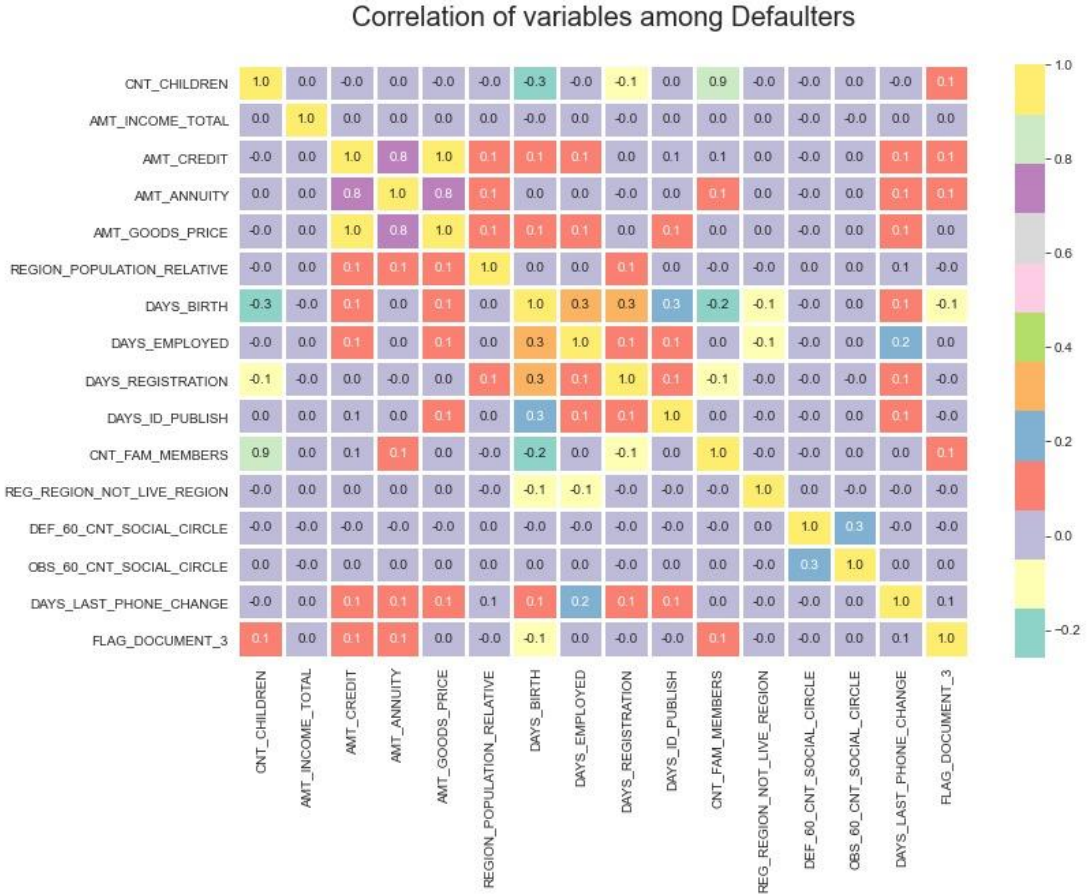
- Lower secondary and Secondary education have the lowest income.

Correlation Matrix

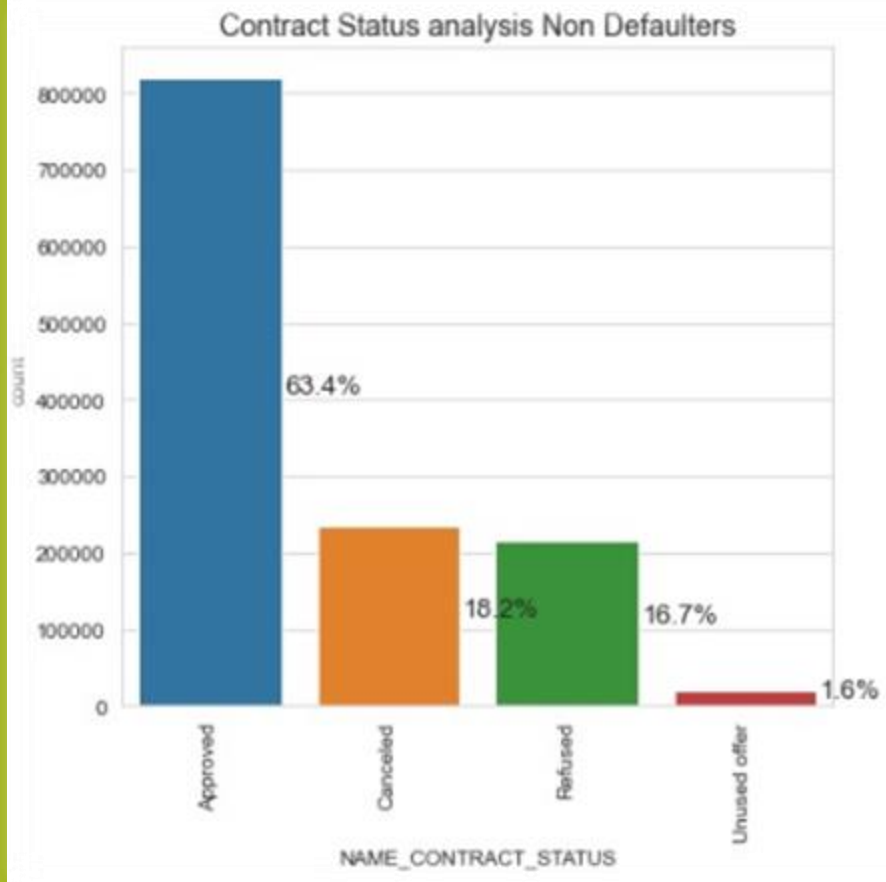
The variables with positive correlation (above 0.4) among both defaulters and non-defaulters are:

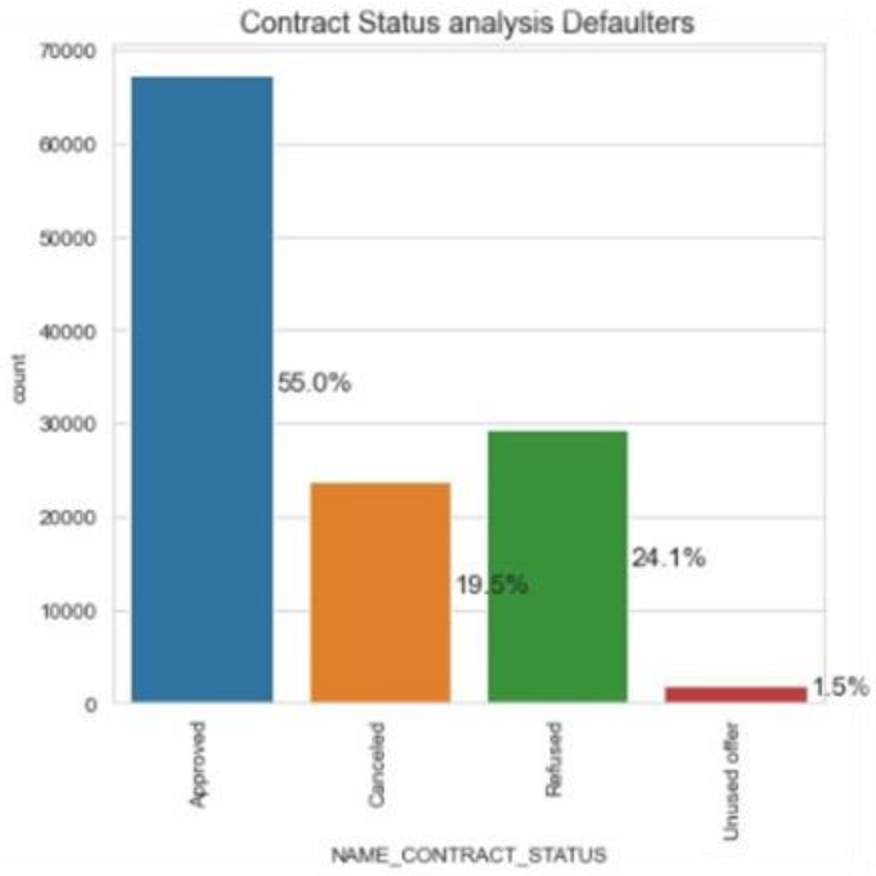
- AMT_CREDIT - AMT_GOODS_PRICE
- AMT_ANNUITY -AMT_INCOME_TOTAL
- AMT_ANNUITY - AMT_GOODS_PRICE
- AMT_ANNUITY - AMT_CREDIT
- CNT_CHILDREN - CNT_FAM_MEMBERS





Merged Analysis
CONTRACT STATUS





Inference:

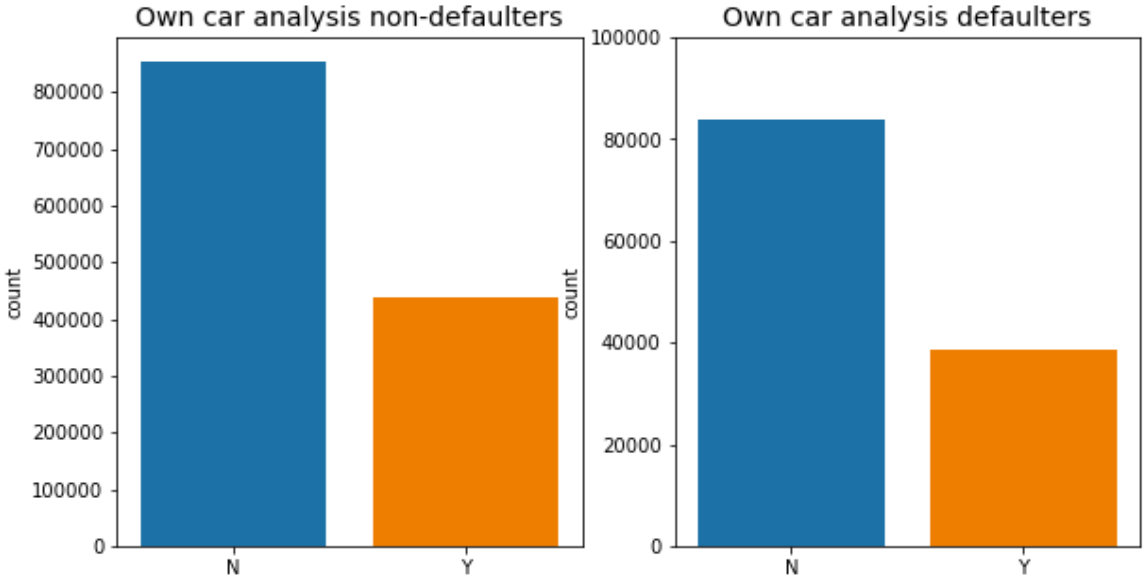
- 90% of applicants who previously had cancelled loans are non-defaulters.

- Applicants who had previously been refused loans are the second highest percentage among defaulters.

- 88% of applicants who had previously been refused loans are non-defaulters.

Merged Analysis

OWN CAR, OWN REALTY

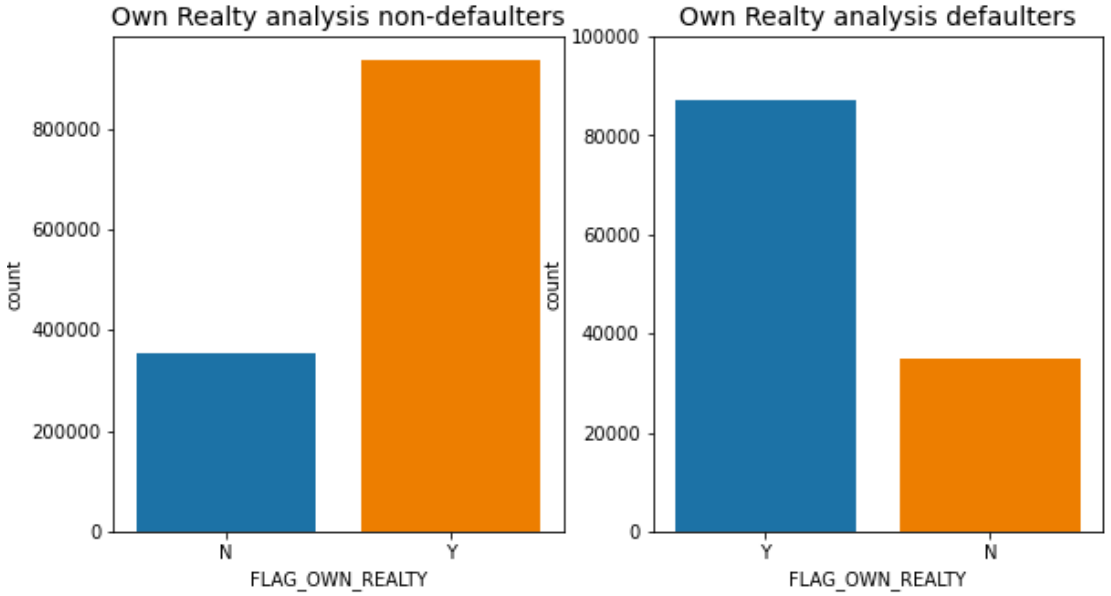


Inference:

- Most applicants do not own a car.
-Applicants without cars have a higher default pecentage

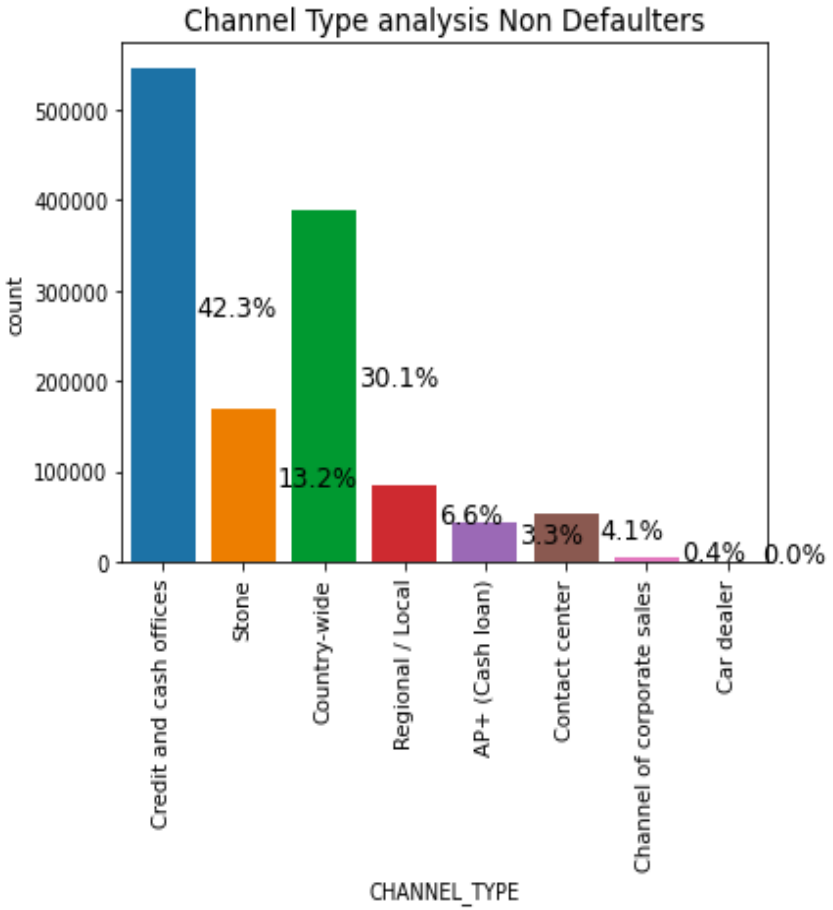
Inference:

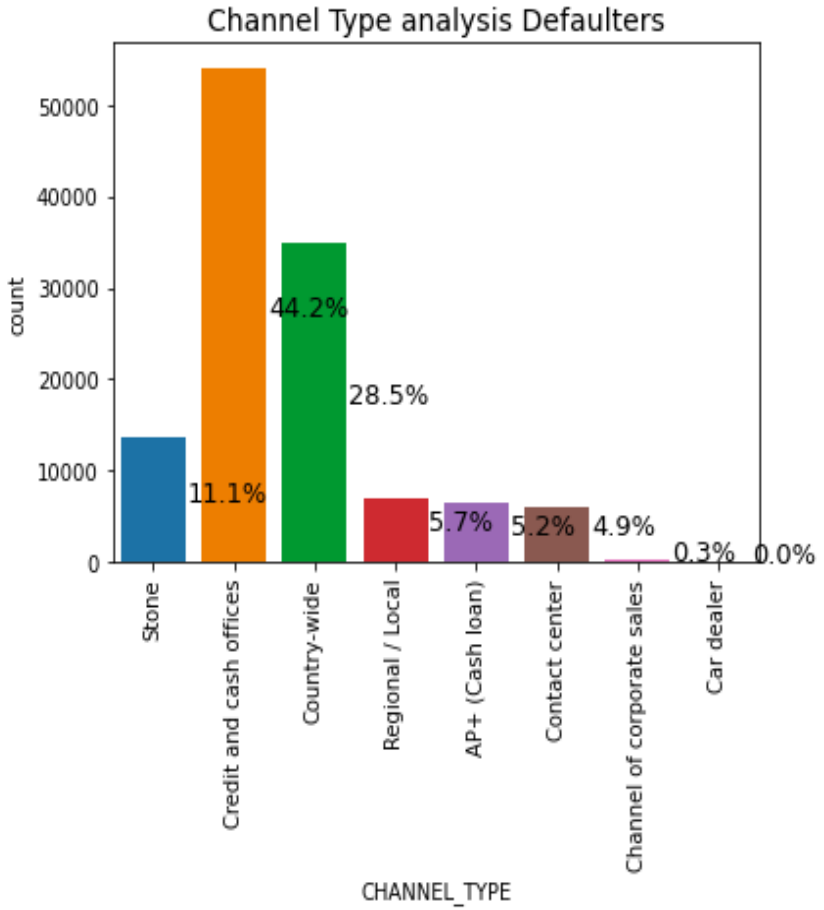
-Applicants that own realty are more likely to be non-defaulters.
- Owning Realty is a high percentage for all applicanants.
- Applicants without own realty have a higher default percentage.



Merged Analysis

CHANNEL TYPE



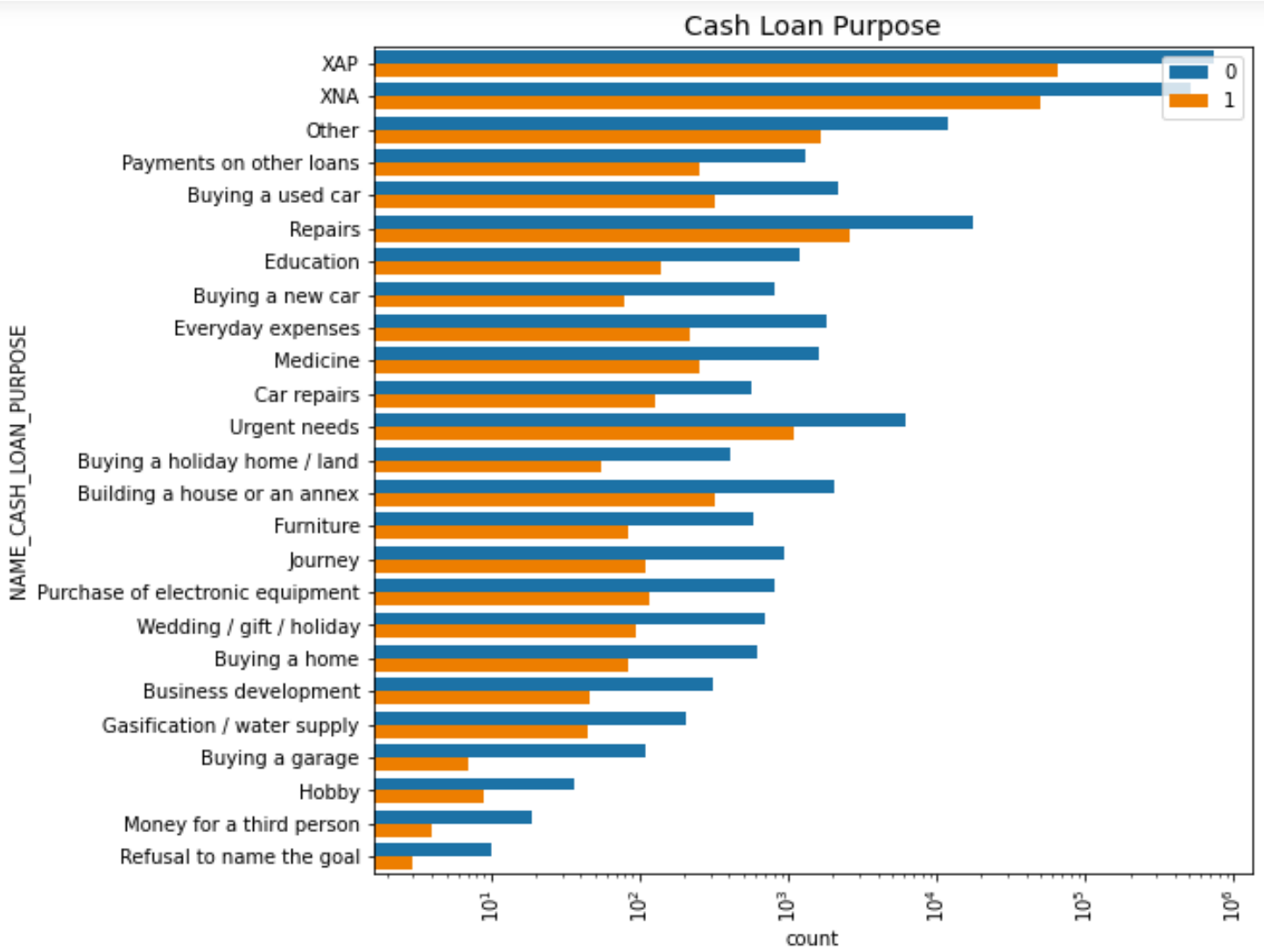


Inference:

- Most loan applicants have previously been acquired through 'Credit and Cash office' and 'Country Wide' channel types.
- Applicants who have previously been acquired through 'AP+(cash loans)' and 'Contact Center' channel type have a higher percentage of Defaulters.
- Applicants who have previously been acquired through 'Car dealer' and 'Channel of corporate sales' channel type have a higher percentage of Non defaulters.

Merged Analysis

CASH LOAN PURPOSE

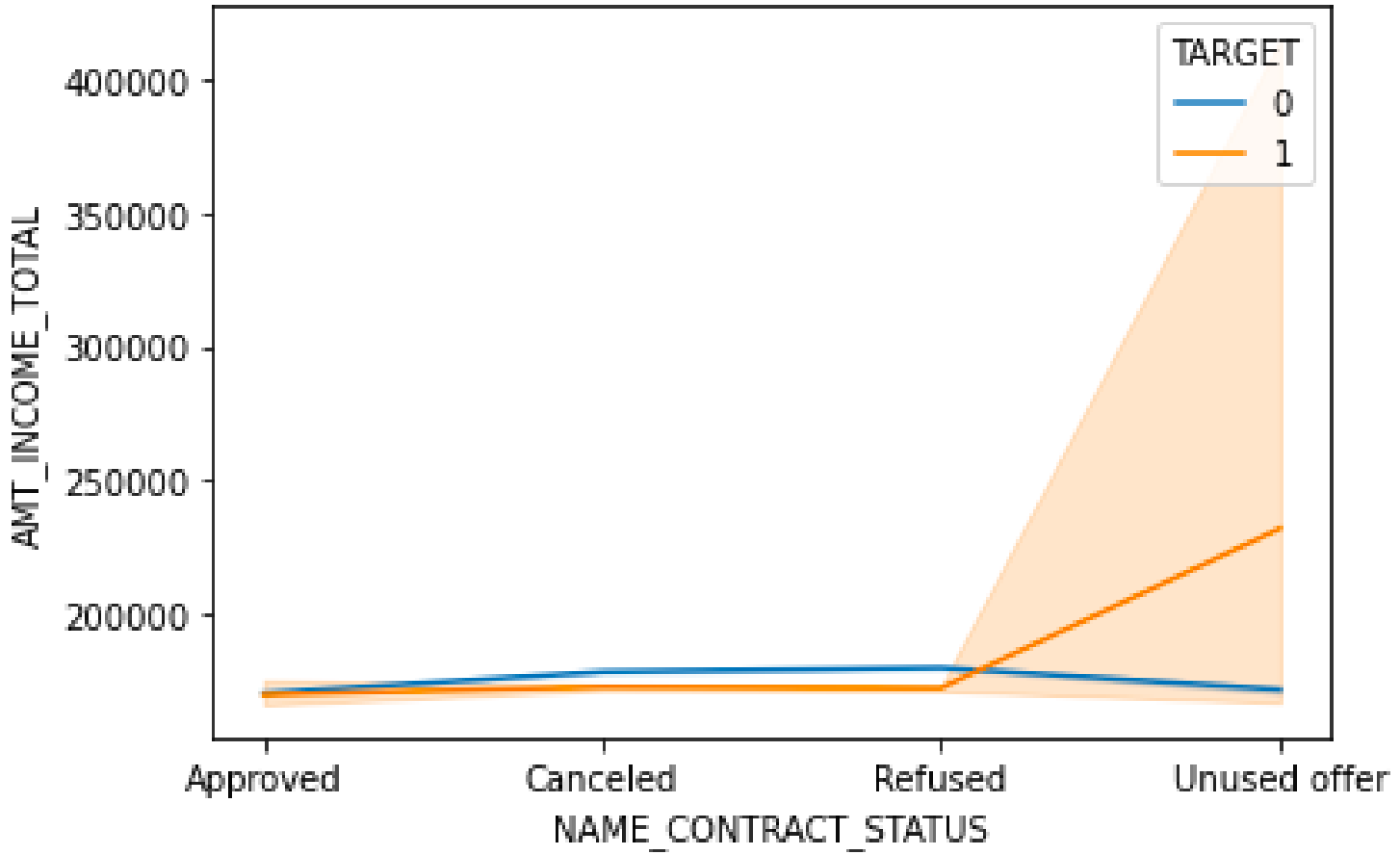


Inference:

-Applicants whose Cash Loan Purpose is Car repairs, and Urgent needs have higher default percentage.

- Applicants with purpose as Buying a new garage, Buying new car and Education have lowest deafult percentage.

Merged Analysis
Income Total



Inference:

- Applicants who have previously unused offers, have defaulted on their loans.

- These defaulter applicants with unused offers, also have relatively higher income total than the average.

Loan Attributes

Non-Defaulters:

Target:

- Gender: Female applicants
- Family Status: Widows and married status
- Occupation: Accountants, high skill tech staff, managers, Students and Businessmen.
- Education: Higher education and academic degree holders.
- Income type: Pensioners, State servants
- Housing Type: Own house/apartment, Office and co-op apartments.
- Age: Over the age of 50.
- Children: with no children, or less than 2.
- External credit score: over 0.5
- Loan purpose: buying new car, garage and education
- Channel Type: 'Car dealer' and 'Channel of corporate sales'

Defaulters

Risky applicants:

- Gender: Male clients
- Family Status: Civil marriage and Single/Not married status.
- Occupation:Low skill laborers,Laborers, drivers, waiters/bar staff and Security staff.
- Education: Lower secondary and Secondary special.
- Income type: Working, Commercial associates.
- Housing Type: Rented apartment, living with parents.
- Age: 30-40 age group
- Children: over 4 or more children
- Family members: 4 or more family members
- External credit score: less than 0.4
- Loan purpose: Car repairs, urgent need
- Channel Type: AP+ Cash Loans and 'Contact Center' channel type

Suggestions

- Create an evaluation rubric for loan applicants based on the loan attributes that are strong indicators for defaulters; with a risk level attached to each application.
- Have stricter evaluations for applicants who meet 3 or more attributes of risky applicants.
- For clients with previously unused offers and higher than average income, inquire reason for not using loan and determine potential risk by comparing financials from the previous application time.
- For applicants with moderate- high risk levels, offer high interest rates and offer partial loans.
- Target clients with loan attributes indicating non-defaulters.Negotiate with these clients when offering loans by offering competitive interest rates.
- Applicants who have previously been refused loans have not defaulted. Inquire reasons for refused loan and introduce re-examination of loans for refused applicants to target them for future loan offers.