CREDIT - EDA Case Study

Partners :

Dadi Padmakara Srinivas

Lakshmi Gayathri MNV.

Overview

Introduction/Problem Statement

Business Understanding

Approach

Data Understanding

Data Cleaning

Data Blending

Univariate and Multivariate Analysis

Insights and Conclusions

Introduction

- This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will 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 Understanding

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

Business Understanding - Types of Risks

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

2. If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Business 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:** All other cases when the payment is paid on time.

Business Understanding

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

Approved: The Company has approved loan Application

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.

Refused: The company had rejected the loan (because the client does not meet their requirements etc.).

Unused offer: Loan has been cancelled by the client but on different stages of the process.

In this case study, we will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Exploratory Data Analysis(EDA) Approach

- Data Understanding
- Data Cleaning
    1. Missing value percentages and Columns Elimination
    2. Sanity Checks
- Data Blending
- Data Visualization
    1. Univariate Analysis
    2 .Multivariate Analysis
- Insights Conclusion

Overall Data Understanding

Dataset Name	Total Rows	Total Columns	Categorical columns count	Numerical Columns count
Application Dataset	307511	122	37(30%)	85(70%)
Previous Application Dataset	1670214	33	18(55%)	15(45%)

- In application dataset Out of 122 columns , 30% are categorical columns and 70% columns are numerical.

  Hence these many columns are not irrelevant, we performed analysis only for few columns.

- In Previous Application Dataset out of 33 columns,54% are categorical columns out of total columns.

Data Cleaning Sanity Checks

- Missing values columns Identification and Elimination
- Eliminate Unnecessary Columns for analysis.
- Convert Negative values into positive values.
- Data Types should be mapped properly in both the datasets.

Data Cleaning – Missing Values Identification & Elimination

Dataset Name	Total Rows	Total Columns	Missing Values columns	Missing Values Percentage Threshold	Elimination Columns	Total Columns considered for Analysis.
Application Dataset	307511	122	64(52%)	40%	49(40%)	73
Previous Application Dataset	1670214	33	14(42%)	50%	4(12%)	29

- In application Dataset, there are 49 columns has highest missing values percentage(90% of the data is missing in 49 columns), so we decided to eliminate them and we cannot impute the values for 49 columns.
- In Previous Application Dataset, there are 4 columns, we eliminated it . We changed the threshold for previous application dataset because there are few important columns for business analysis like DAYS_FIRST_DRAWING etc..!!

Unnecessary Columns Elimination For Analysis in Application Dataset

External sources:

- These columns are irrelevant because it has the information collected from the external data, since we don't know the exact source from this data collected we cannot get any meaningful insights from them.

Flag Documents:

- Flag documents are the various types of documents provided by customer , since we don't know what types of documents provided by the customer, we can eliminate them completely.

Apartment Variables:

- These are the normalized information from the external source. These might not be helpful in analysis with target variable.

Conversion - Negative to Positive values

Application Dataset:

"DAYS_BIRTH","DAYS_EMPLOYED","DAYS_REGISTRATION" and "DAYS_ID_PUBLISH"

- Since Days birth, Days Employed cant be negative , so need to convert them into positive in application Dataset. Similarly this will be applicable for above columns.

Previous application Dataset:

'DAYS_FIRST_DRAWING','DAYS_FIRST_DUE','DAYS_LAST_DUE','DAYS_LAST_DUE_1ST_VERSION','DAYS_TERMINATION','DAYS_DECISION','SELLERPLACE_AREA'.

- Since Days cannot be negative, so we converted into positive values.

Data Types Conversion

Since data types should be mapped properly, it means categorical variable cant be string or object. It should be mapped into category only like Target, Name type suite etc.

Few columns should be converted to integer because they are meaningless if we cannot convert.

Eg: count of children – 2.5 it means 2.5 children is there in a family, it is meaningless and we mapped it into integer. After mapping into integer count of children will be 2,it means family has 2 children.

Data Blending

- Blend the both application dataset and previous application dataset using Current SK ID(Loan ID of the sample).
- There are few column columns which are present in both the dataset like amount goods price, name type suite. For that we used current and previous ,it means current indicates the current column in the application dataset and previous means column in the previous dataset.

Data Blending(Stats) - Merged Dataset

Dataset Name	Total Rows	Total Columns	Missing Columns	Missing Percentage
Merged Dataset(Application and previous dataset)	1430155	74	45	0-40%

- Hence there are 45 columns has missing values but it contains the maximum less than 40% of the missing data in each column. Hence we can proceed to analysis, these columns will not effect analysis.
- Suggestion : For these columns, we can do imputation based on the following criteria.
- Categorical columns - Mode
- Numerical columns – Mean Median

Data Visualization Summary

- Univariate Analysis

  1.Categorical Analysis.

  2. Numerical Analysis.

- Bi-Variate /Multivariate Analysis

  1. Numerical to Numerical Analysis(Scatterplots, line plots)

  2. Categorical to Numerical Analysis

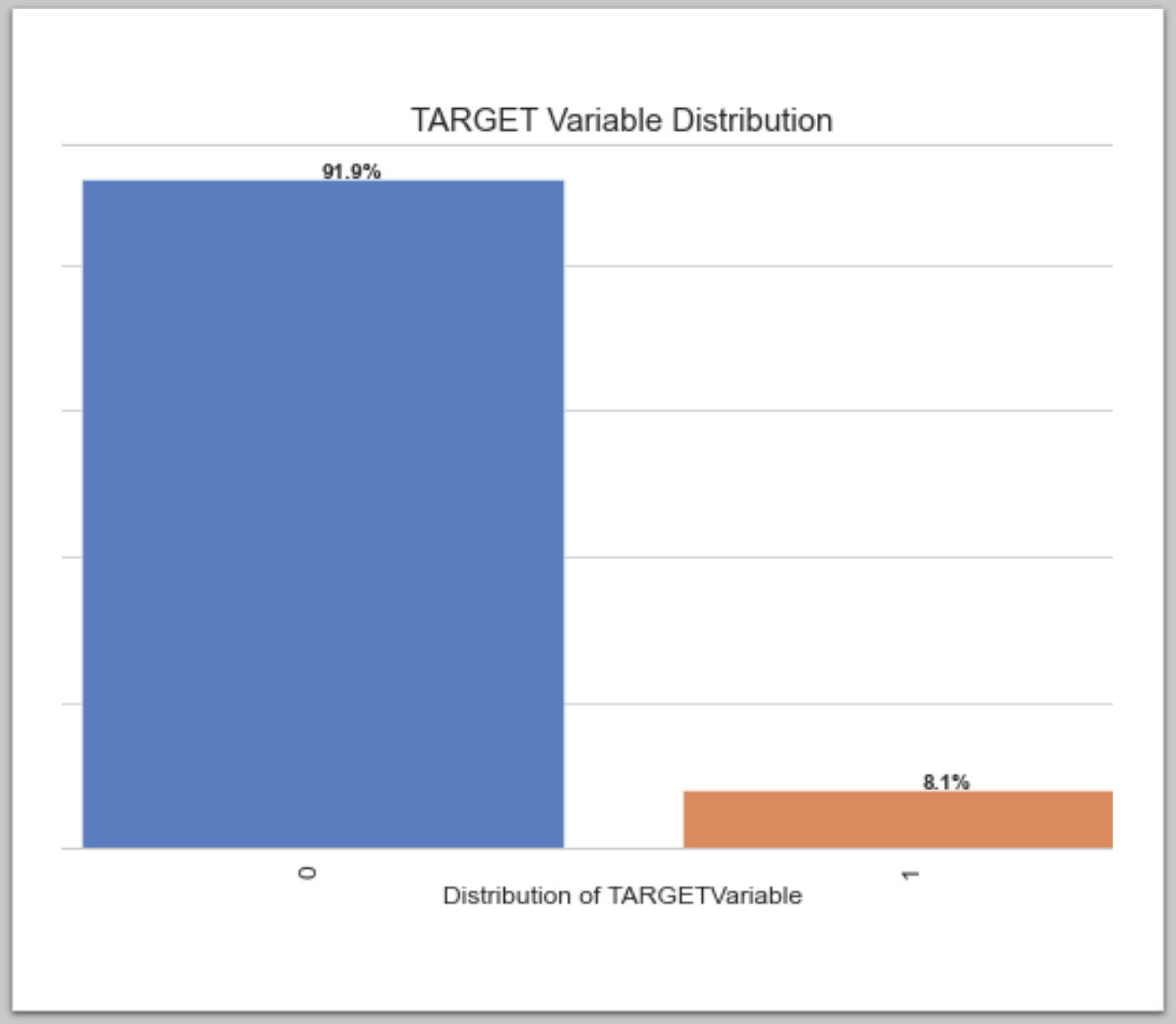
  3. Categorical to Categorical Analysis.

  4. Numerical to Categorical Analysis(More than 2 Categories vs numerical).

Univariate Analysis- Categorical Application Dataset

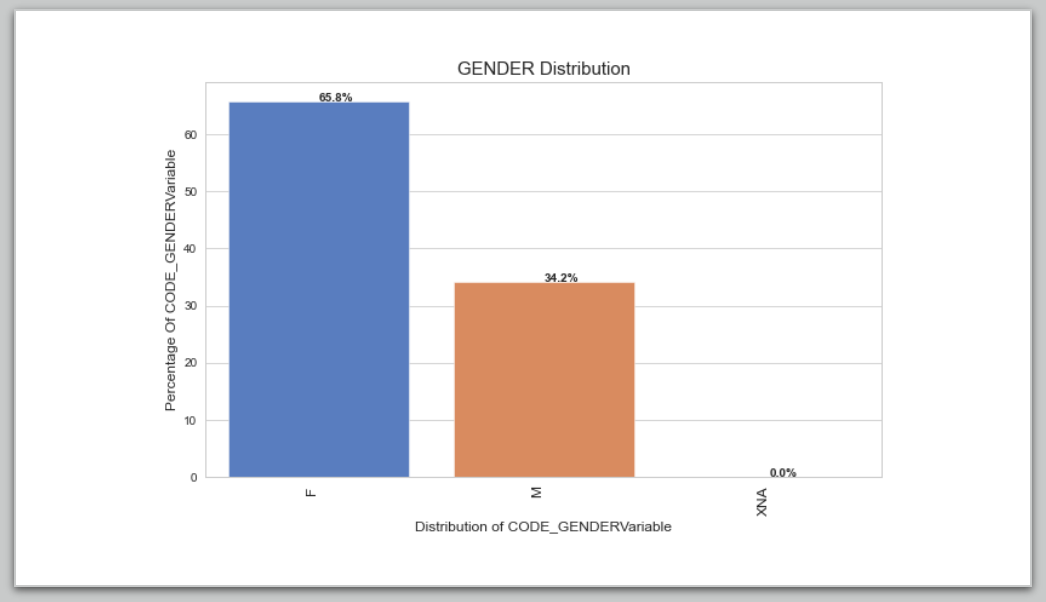
Univariate Analysis : Target Variable

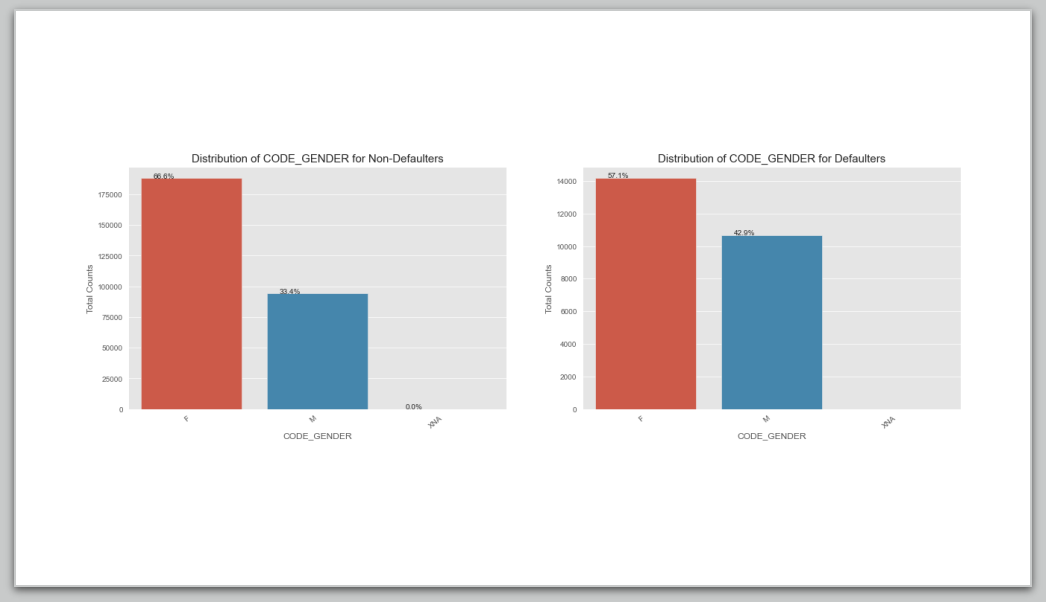
- Target Variable is highly imbalanced, Taget_0 = 91.9% & Target_1 = 8.1%
- Hence the Non- Defaulters contribution is very high & Defaulters contribution is comparatively less in the application dataset.
- By segmenting data into Defaulters &Non Defaulters, we can visualize the reasons/factors for each segment.



Code Gender:

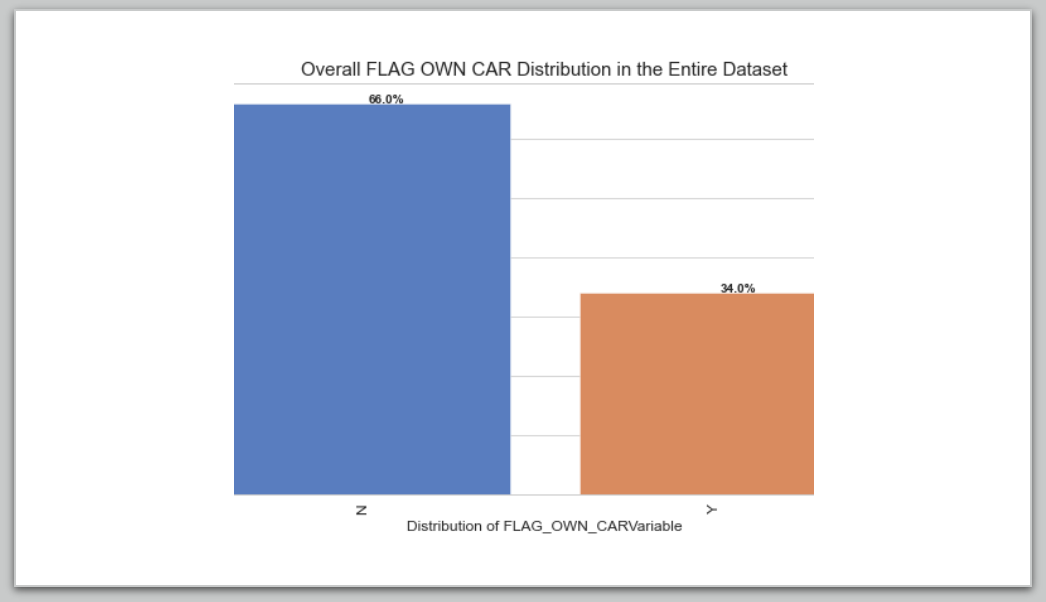
- Overall Gender Distribution in application dataset is , Female contributes 65.8% over Males who contribute only 34.2%
- For Non-Defaulters, Females appear to have a majority of 66.6% over Males with 33.4%.
- For Defaulters, Females appear to be with a majority of 57.1% over Male with 42.9%.
- Hence Female Gender appears to be the highest defaulters in the Application Dataset.

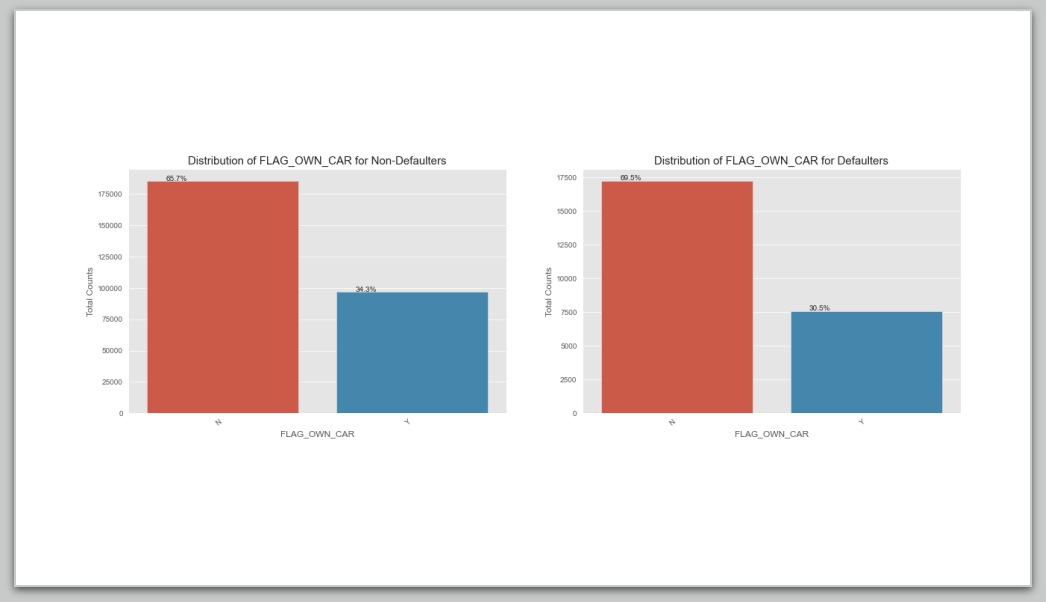




FLAG_OWN_CAR:

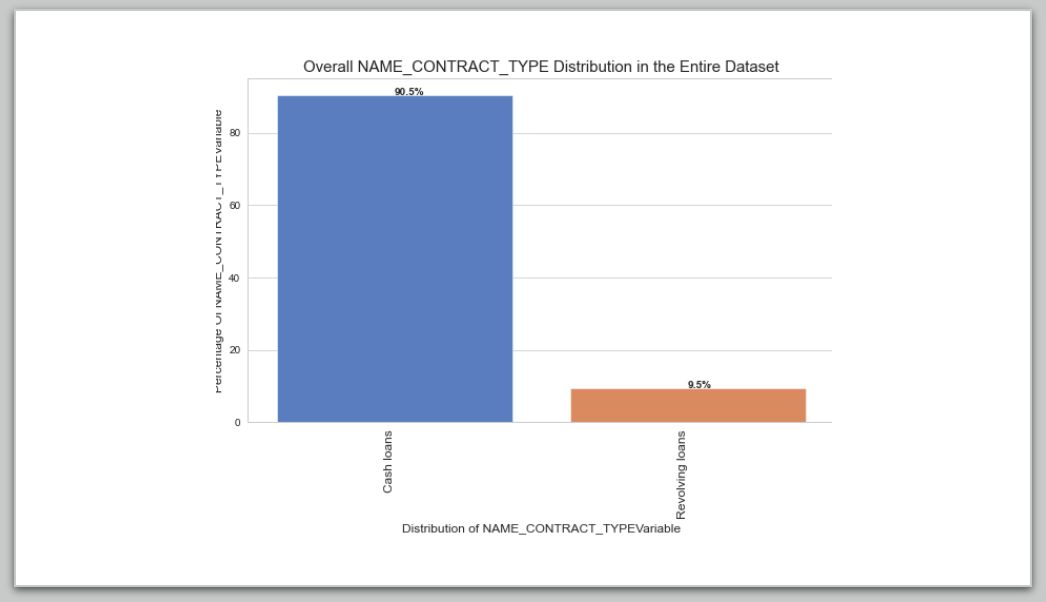
- In application dataset almost 66.0% of the applicants don't have a car & 34.0% have a car.
- For Non-Defaulters almost 65.7% don't have Car & 34.3% have a Car.
- For Defaulters, almost 69.5% don't have a car and 30.5% have a car.
- Hence the most Defaulters seems to have no Car.

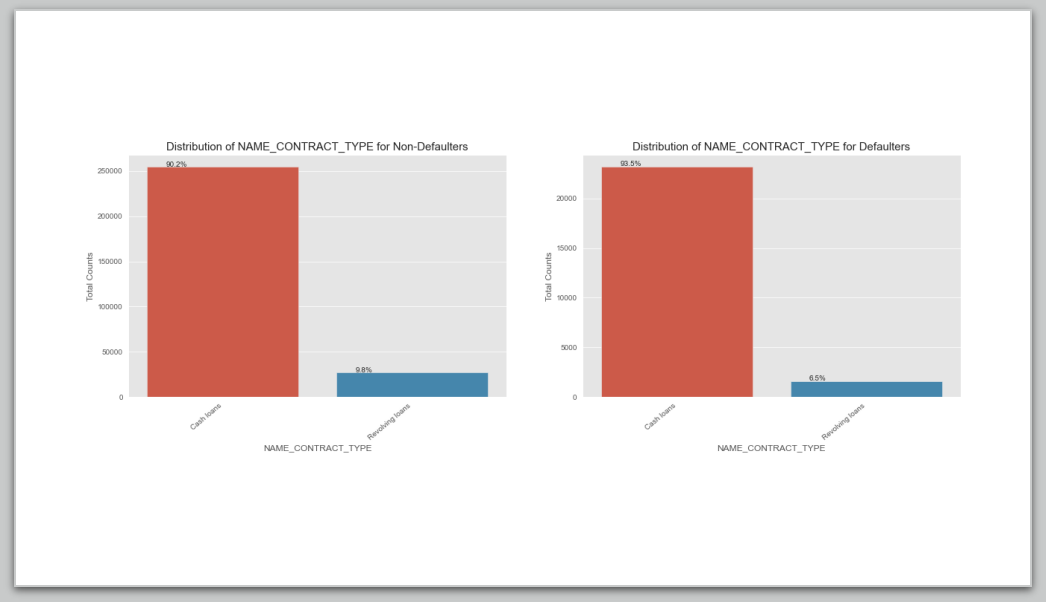




NAME_CONTRACT_TYPE

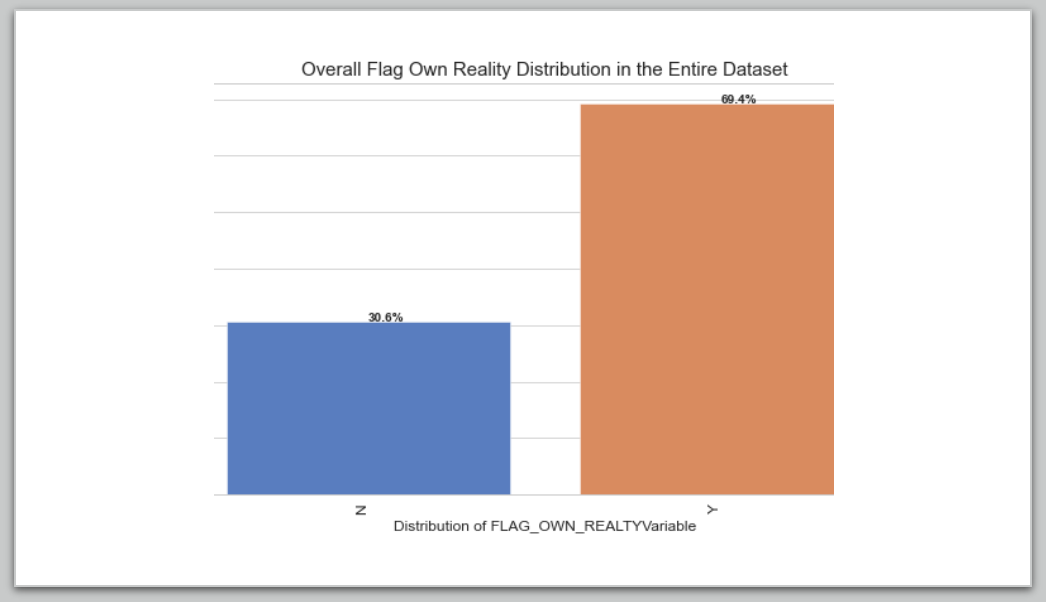
- In Application dataset, almost 90.5% of the applicants have taken Cash Loans & only 9.5 % have taken Revolving Loans.
- For Defaulters, Cash Loans is 90.2% & Revolving Loans is 9.8 %.
- For Non-Defaulters, Cash Loans is 93.5% && Revolving Loans is 6.5%.
- Hence Cash Loans appear to be the highest defaulters Contract Type.

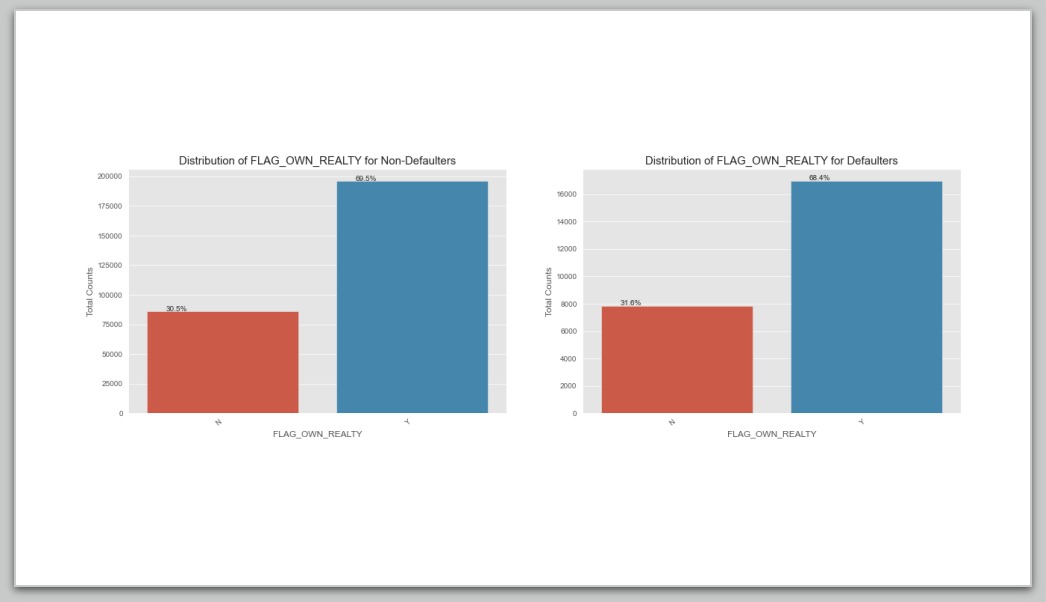




Flag Own Reality:

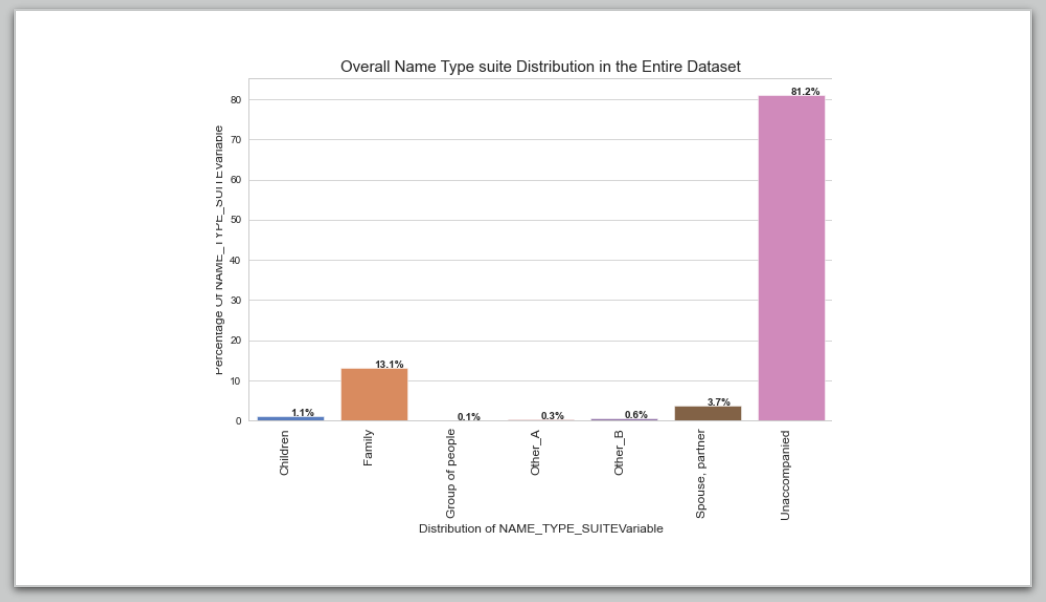
- In the application Dataset, majority of the clients who own Reality as an Reality is 69.4% & who do not own a Reality is 30.6%.
- For Non-Defaulters, who own Reality is 69.5% to that of who do not own a Reality is 30.5%.
- For Defaulters, who own Reality is 68.4% over to those who do not own a Reality is 31.6%.
- Hence Majority of the Defaulters own a Reality.

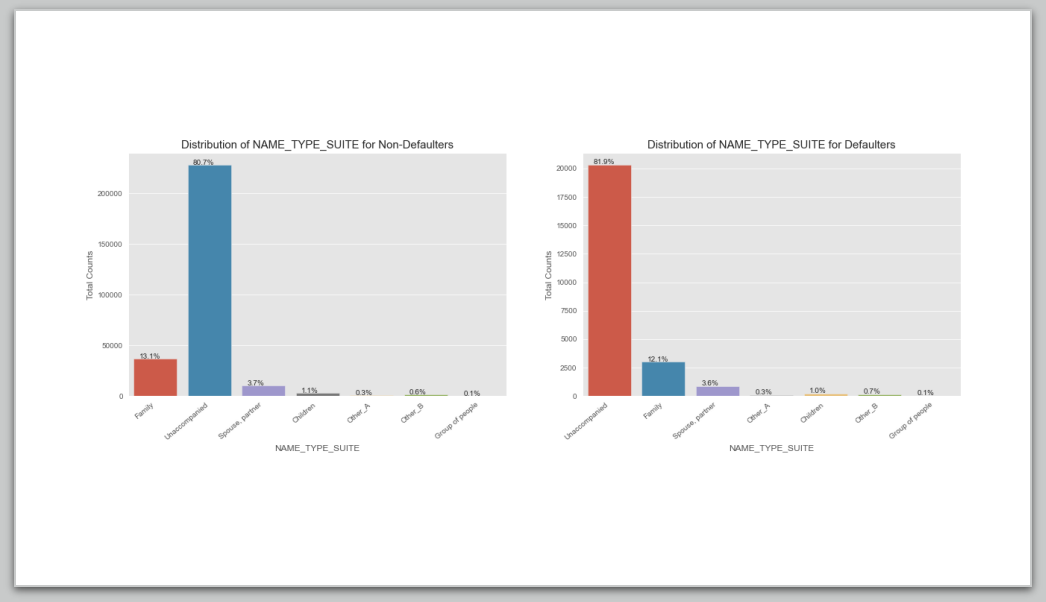




NAME_TYPE_SUITE:

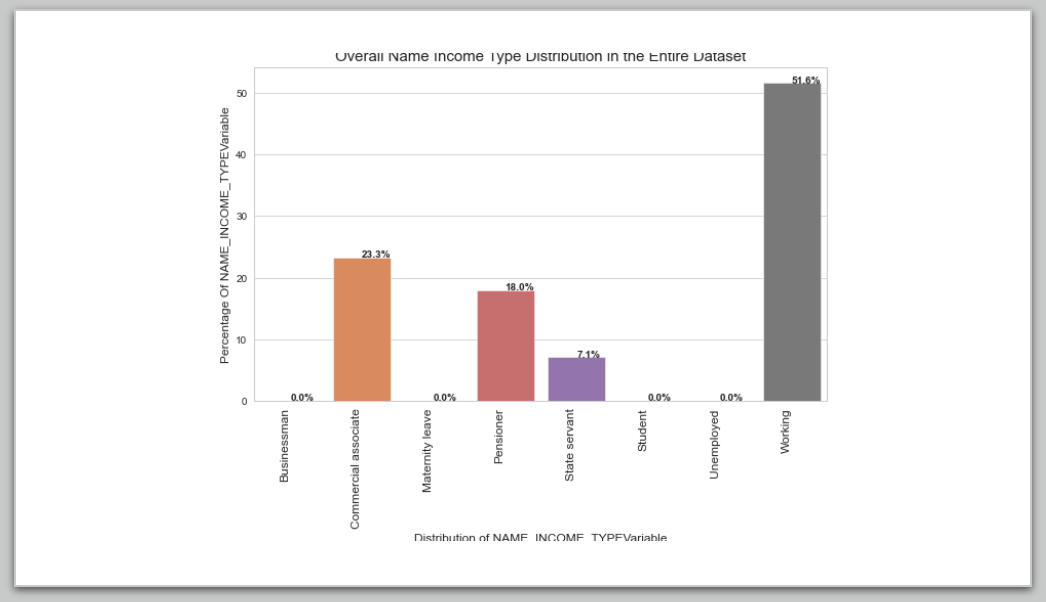
- Unaccompanied appears to be the highest with 81.2 % of clients & Group of People appears to be the least with 0.1% type of Suite in the application dataset.
- Majority of the Non Defaulters are Unaccompanied with 80.7% and least is Group of People with 0.1%.
- Hence, Majority of the Defaulters are also Unaccompanied with 81.9% and the least is Group of People with 0.1%.

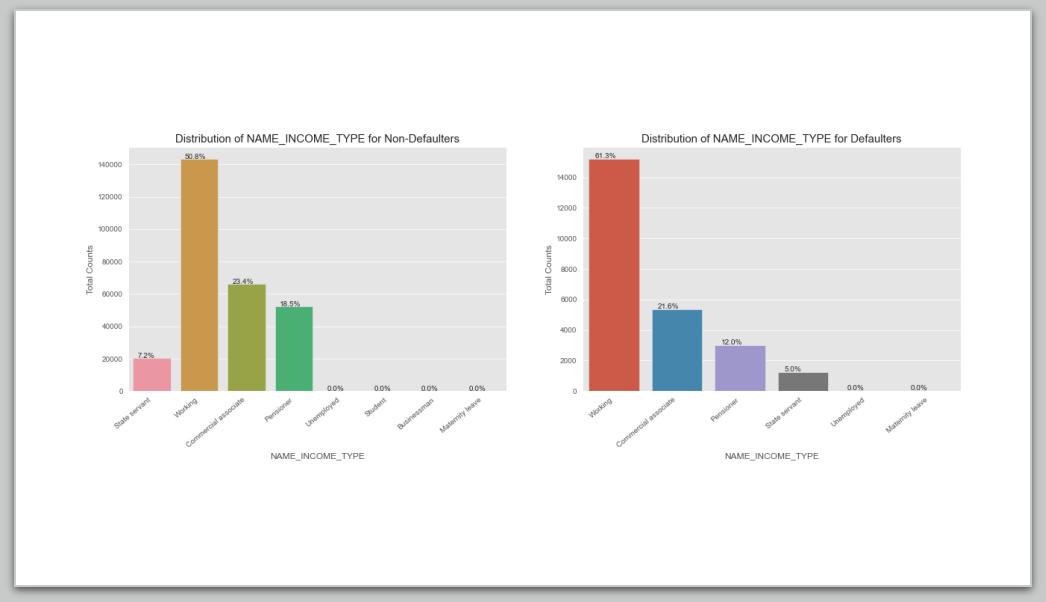




Name Income Type

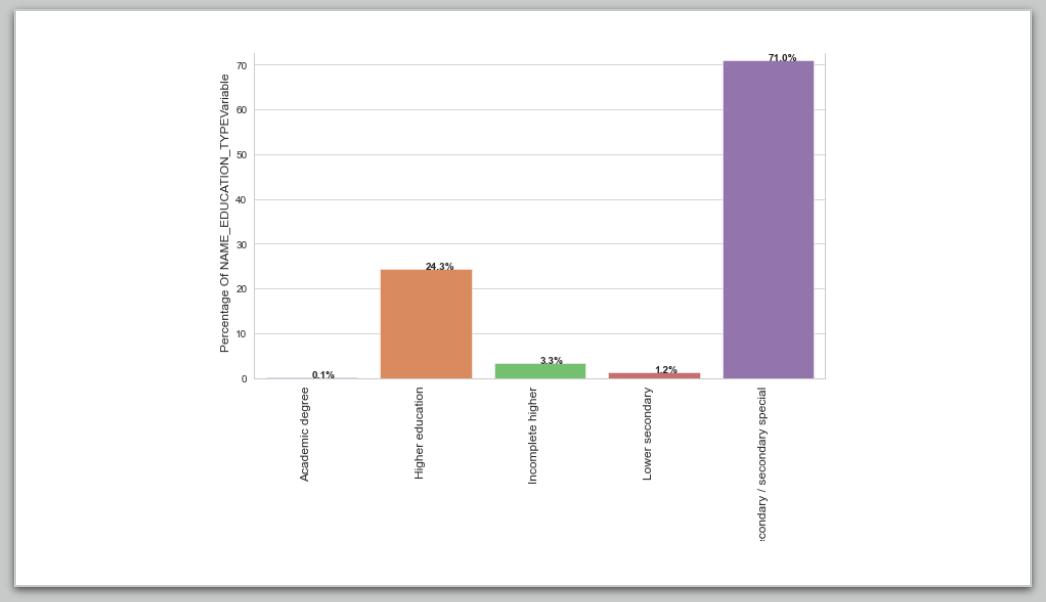
- Working Income Type appear to be the majority in application dataset with 51.6% & State Servant appears to be least with 7.1%.
- Working Income type appears to be majority with 50.8% & State servant is least with 7.2% among Non-Defaulters.
- Working Income Type appears to be majority with 61.3% & State Servants appear to be 5.0% among Defaulters.

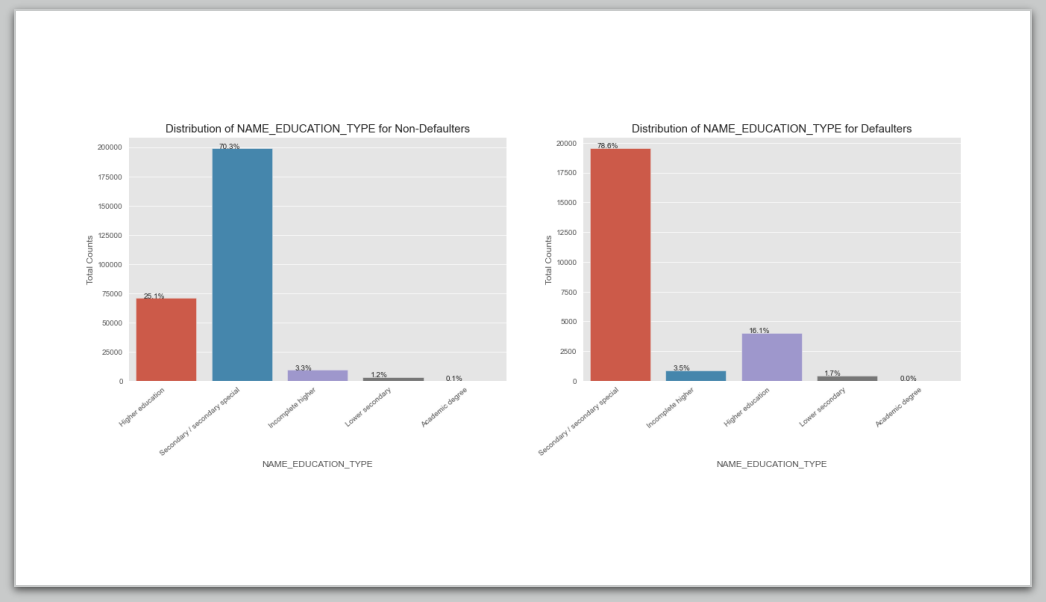




NAME Education Type:

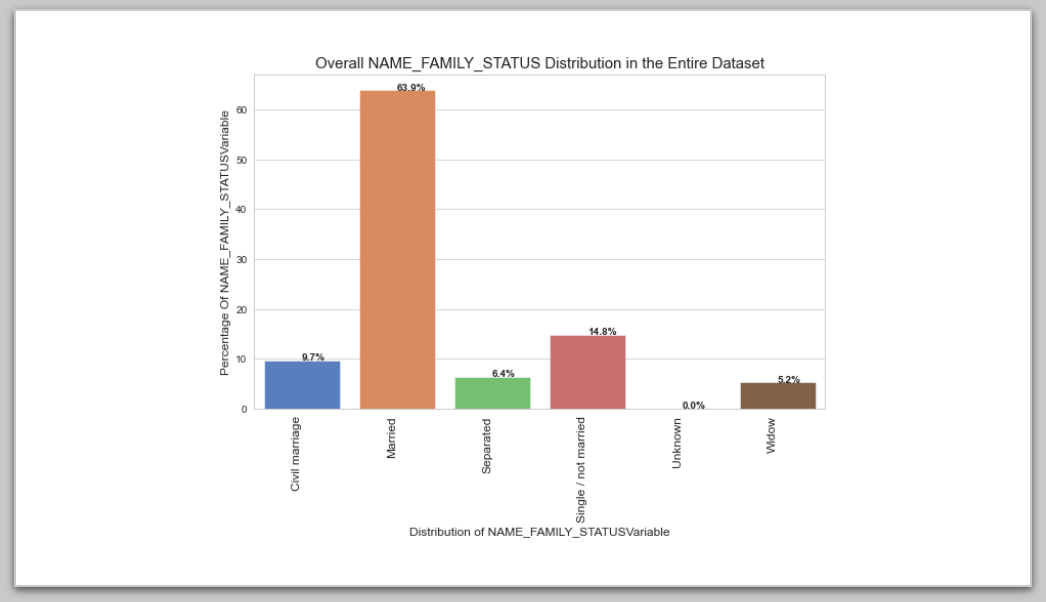
- Secondary/Secondary Special appears to be majority of the education of the clients with 71.0% & Academic degree appears to be the least with 0.1% of the entire clients in the application dataset.
- In Non- Defaulters, Secondary/Secondary Special appears to be majority among the education types with 70.3% & Academic degree appears to be the least with 0.1%.
- In Defaulters, Secondary/Secondary Special appears to be majority among the education types with 78.6% & Lower Secondary appears to be the least with 0.1%.

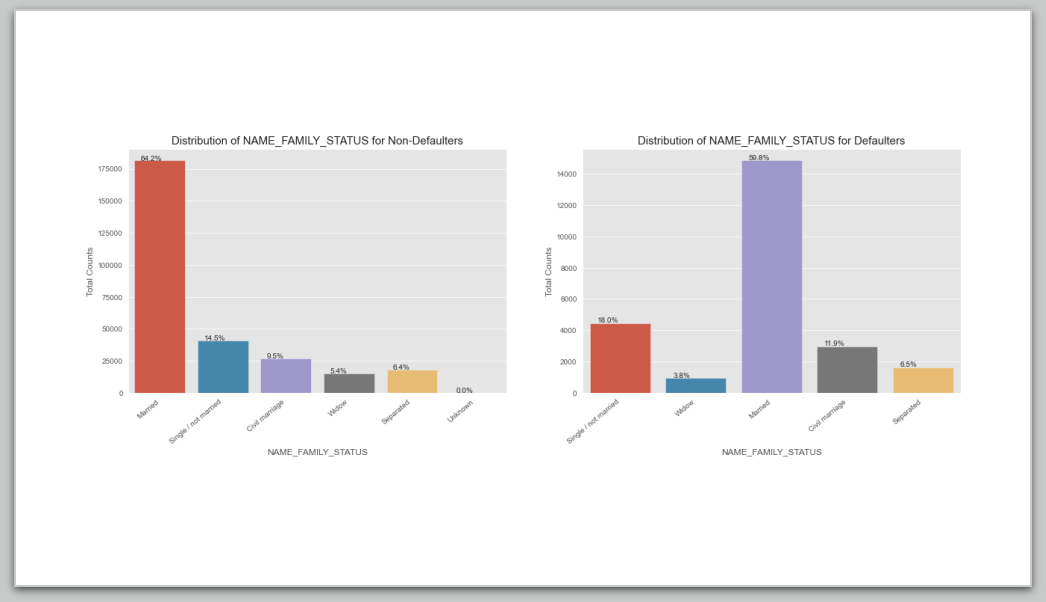




NAME_FAMILY_STATUS

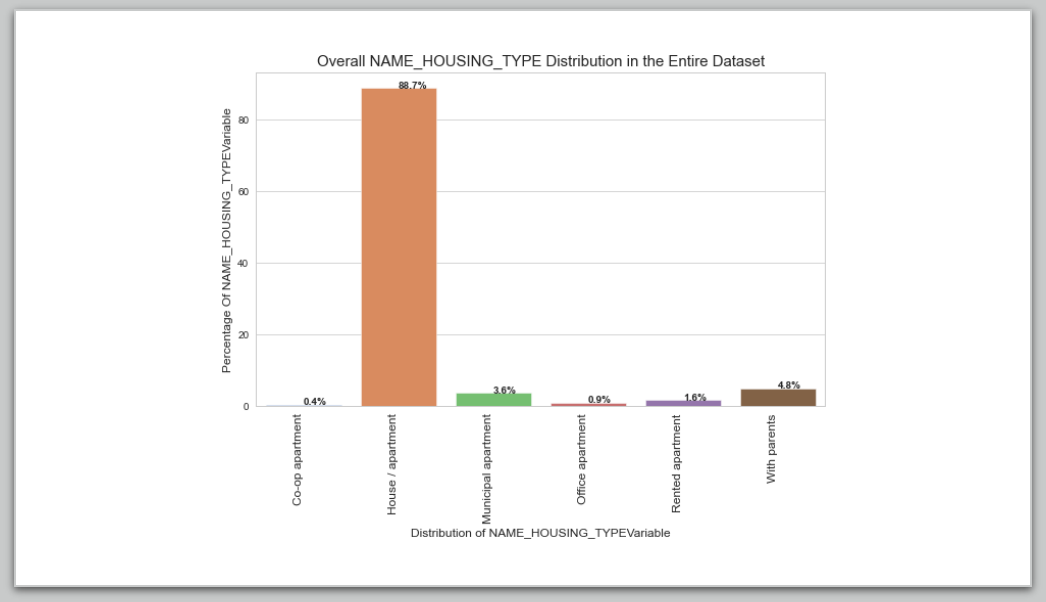
- Married clients appear to be the majority with 63.9% & least appears to be Widow with 5.2% among the Family Status types in the application dataset.
- Among Non-Defaulters, Married Clients are the majority with 64.2% and the least type is Widow with 5.4%.
- Majority of the Defaulters are Married with 59.8% & minority of the defaulters is Widow with 3.8%.

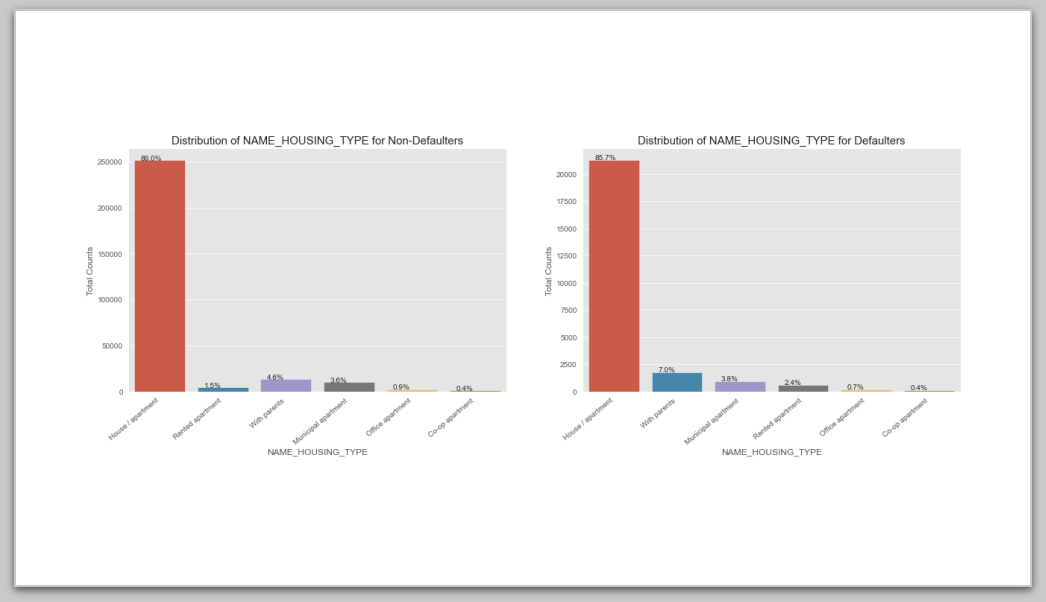




NAME_HOUSING_TYPE

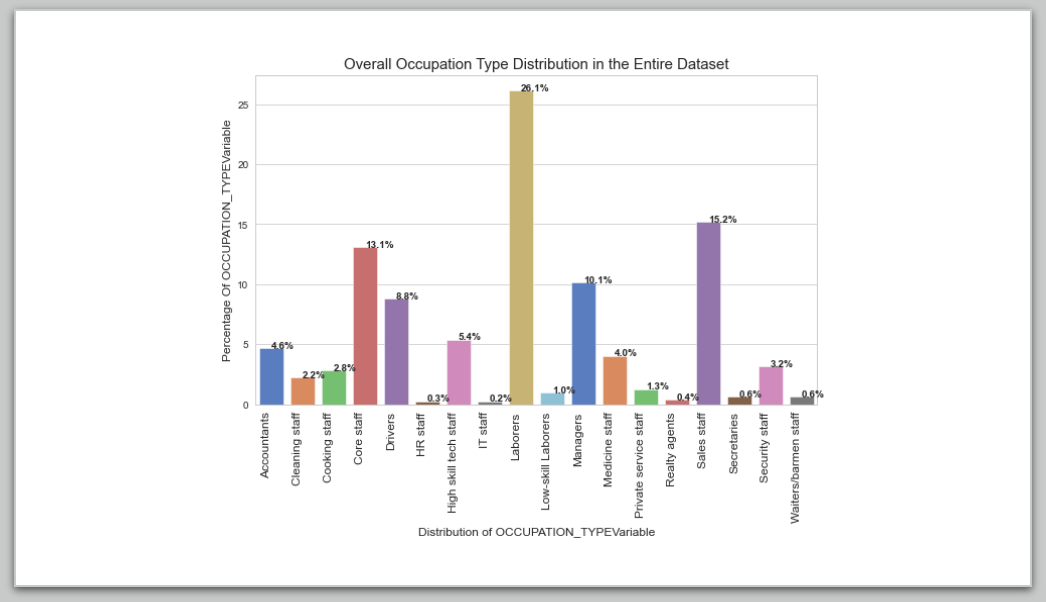
- House/Apartment appears to be majorly owned by 88.7% and least clients own Co-op Apartments with 0.4% by the clients in the application dataset.
- Majority of the Non-Defaulters own House/Apartment with 89.0% and least own Co-op Apartments with 0.4%.
- Majority of the Defaulters own House/Apartment with 85.7% and least own Co-op Apartments with 0.4%.

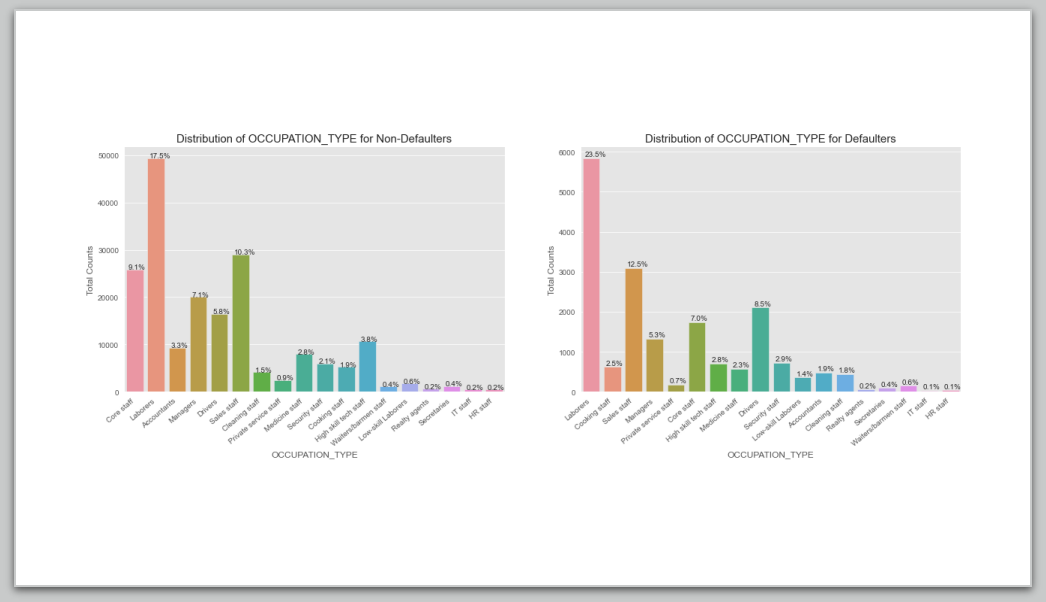




Occupation Type

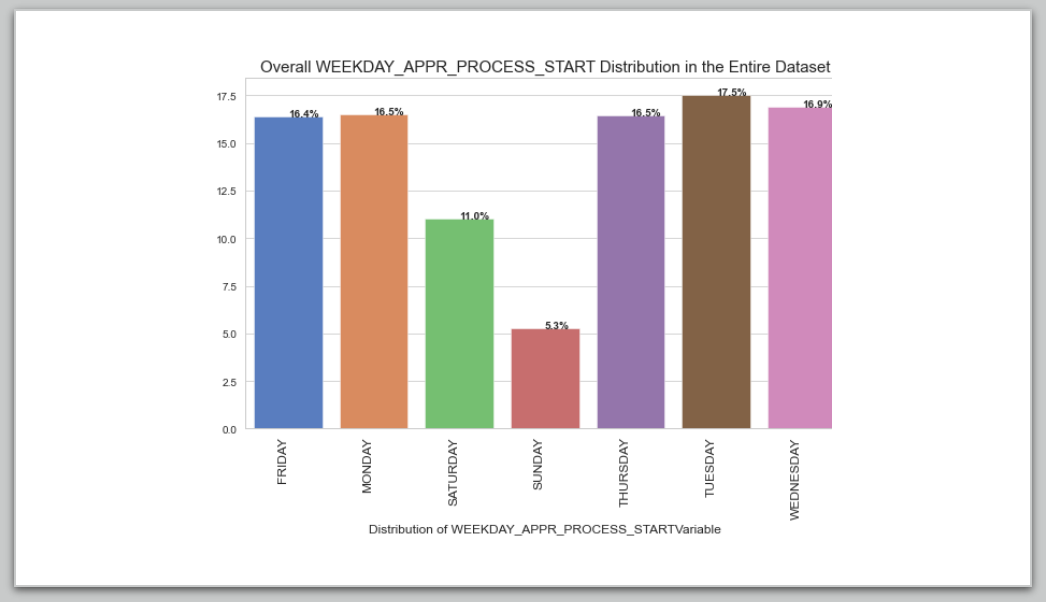
- Majority of the clients are Laborers which constitute to 26.1% and the least is IT Staff with 0.2% among all the clients in application dataset.
- Majority of the Non- Defaulters is Laborers with 17.5% and minority is Reality agents, IT Staff and HR Staff who share 0.2% each.
- Majority of the Defaulters is Laborers 23.5% and least is IT Staff & HR Staff is 0.1%.

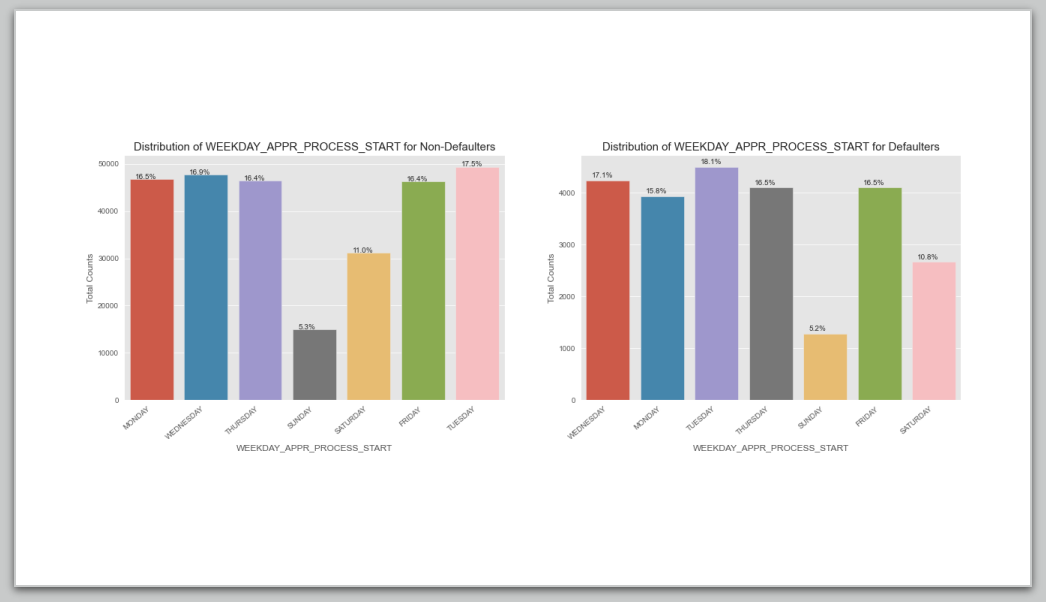




WEEKDAY APPR PROCESS START

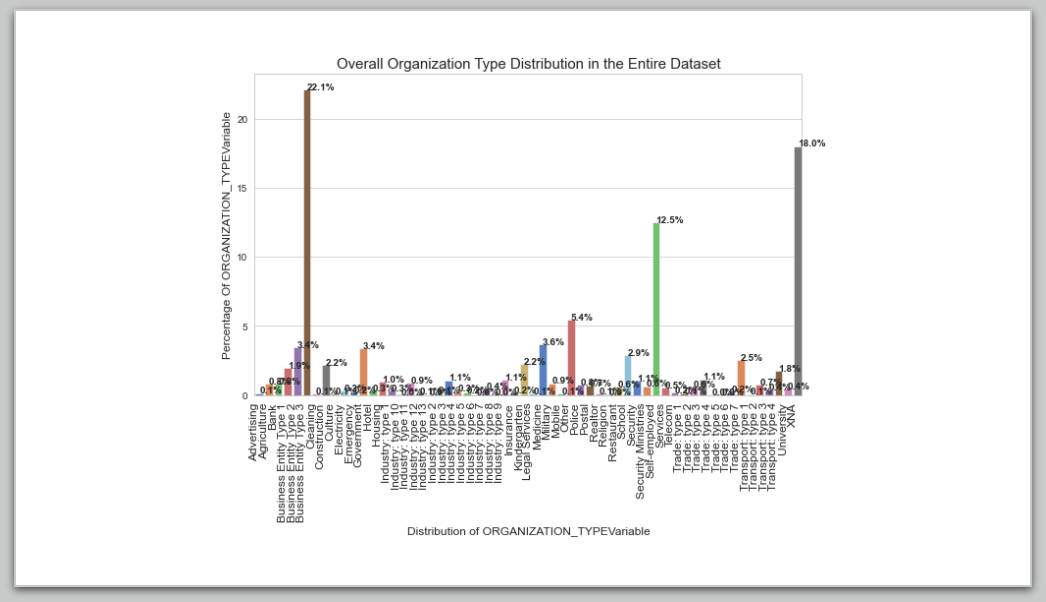
- Majority of the Credit approval start day is Tuesday with 17.5% & least was observed on Sunday with 5.3%.
- Most of the Non-Defaulters Credit approval start day observed to be Tuesday with 17.5% and least was observed on Sunday with 5.3%.
- Most of the Defaulters Credit approval start day is observed to be Tuesday with 18.1% and least on Sunday with 5.2%.

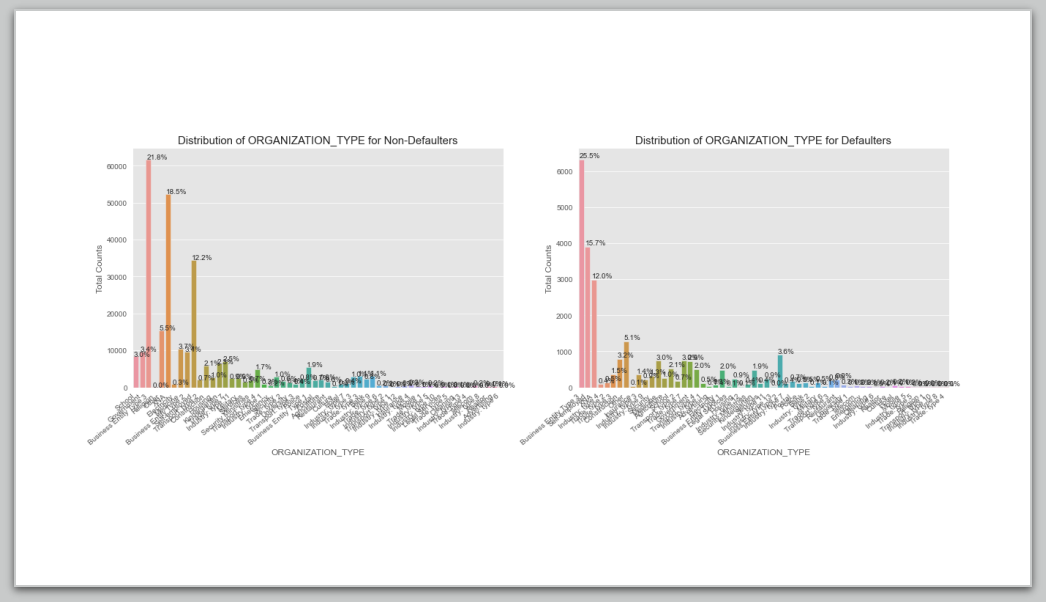




ORGANIZATION_TYPE

- Overall majority of the applicants is Business Entity Type 3 & constitute 22.1%.
- Majority of the Non-Defaulters is also Business Entity Type 3 & constitute 21.8%
- Majority of the Defaulters is also Business Entity Type 3 & constitutes 25.5%.

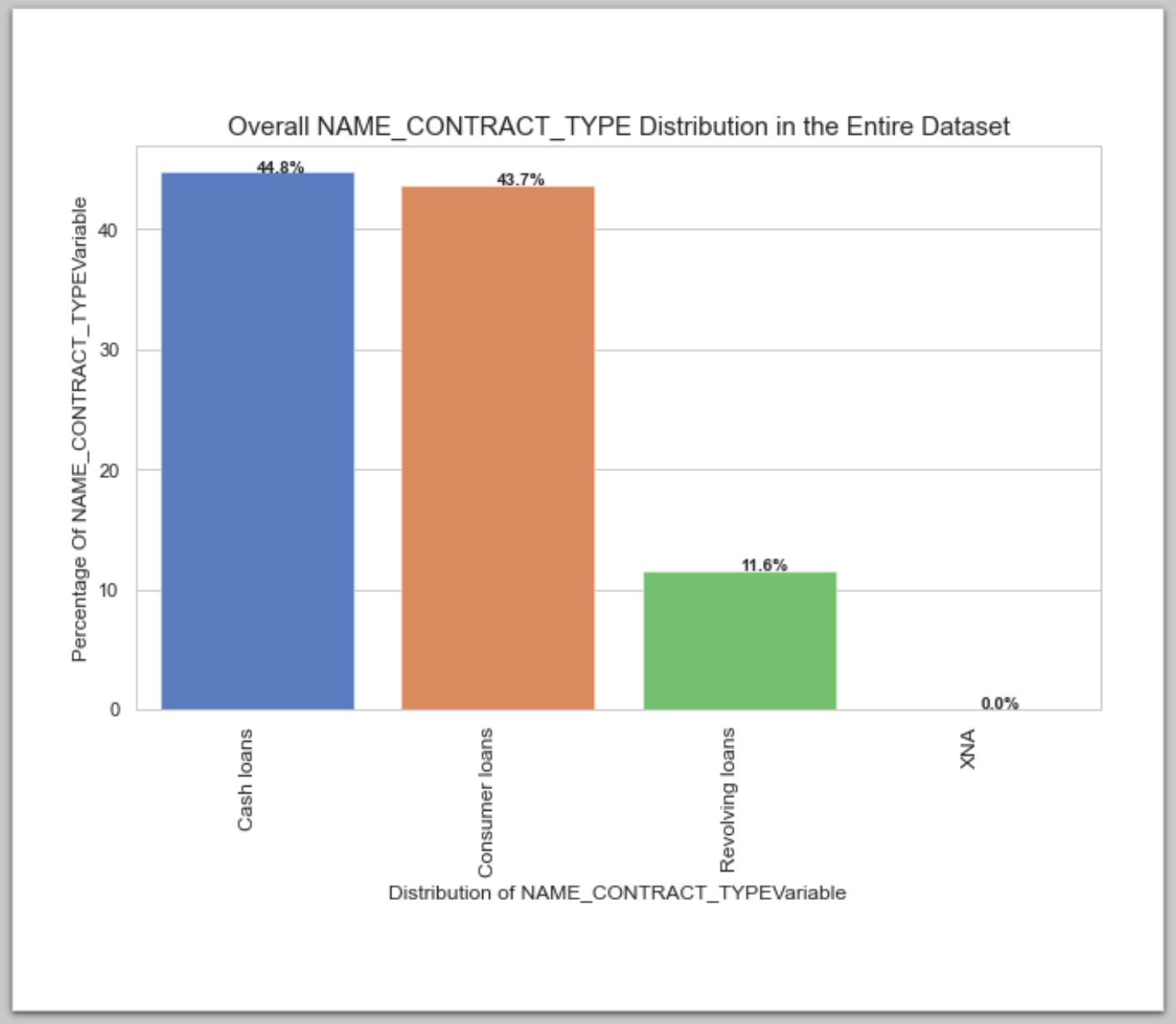




Univariate Analysis – Categorical- Previous Application and Merged Dataset.

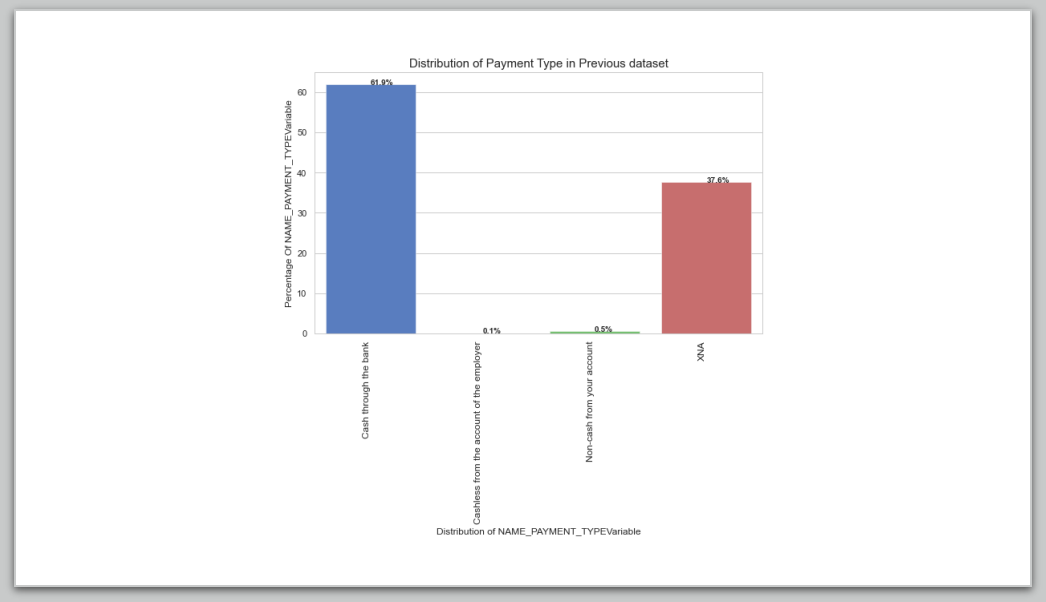
Name Contract Type

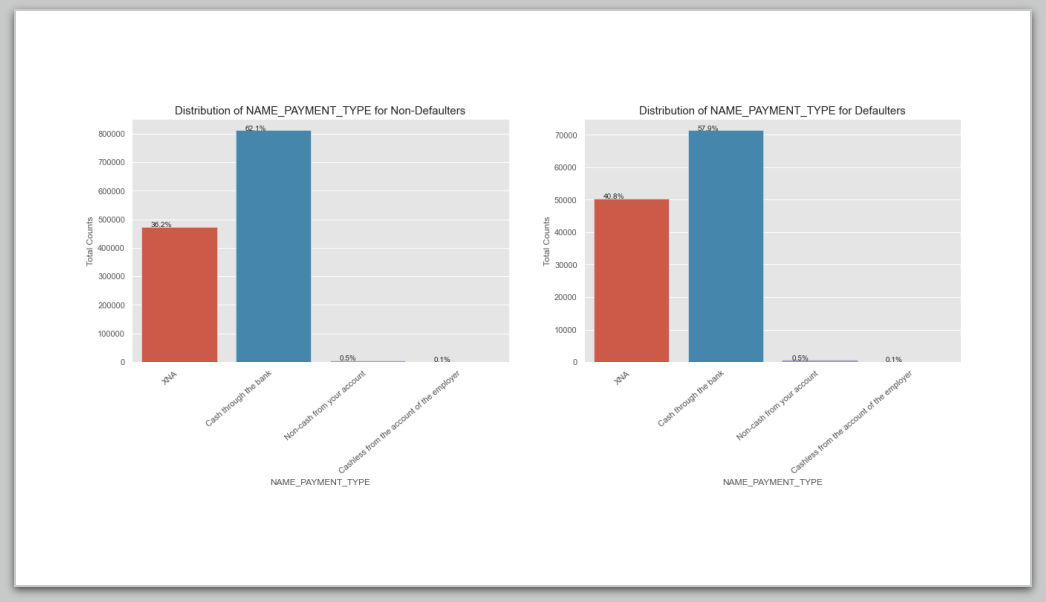
- Cash Loans & Consumer Loans appear to be the maximum type of Loans taken 44.8% & 43.7% of the total applicants.
- Revolving Loans are least taken at rate of 11.6%



Name Payment Type

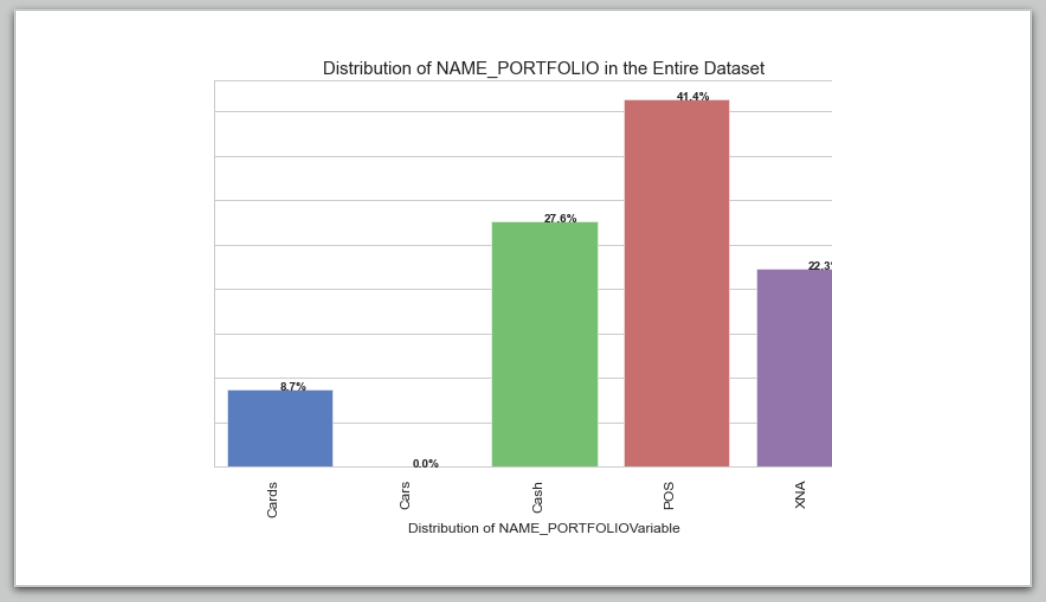
- 61.9% of the Previous applicants opted for Cash through the Bank & least is 0.1% Cashless from the account of the employer.
- 62.1% of the Non-Defaulters have also opted for Cash through the Bank and least is 0.1% Cashless from the account of the employer.
- 57.9% of the Defaulters have also opted for Cash Through the Bank and least is 0.1% Cashless from the account of the employer.
- So majority of the applicants opted for Cash through the Bank.

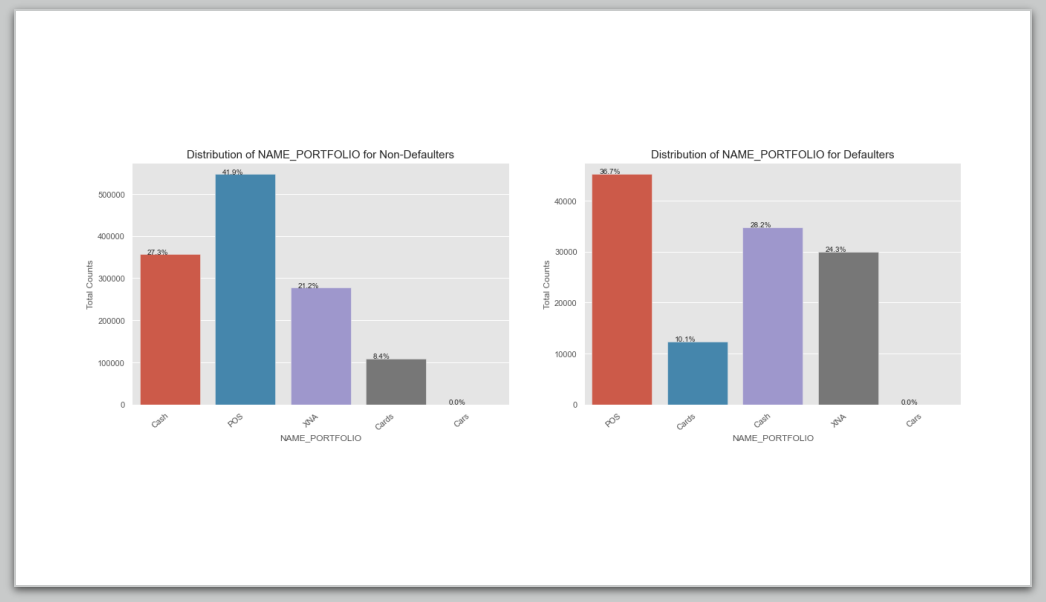




NAME_PORTFOLIO

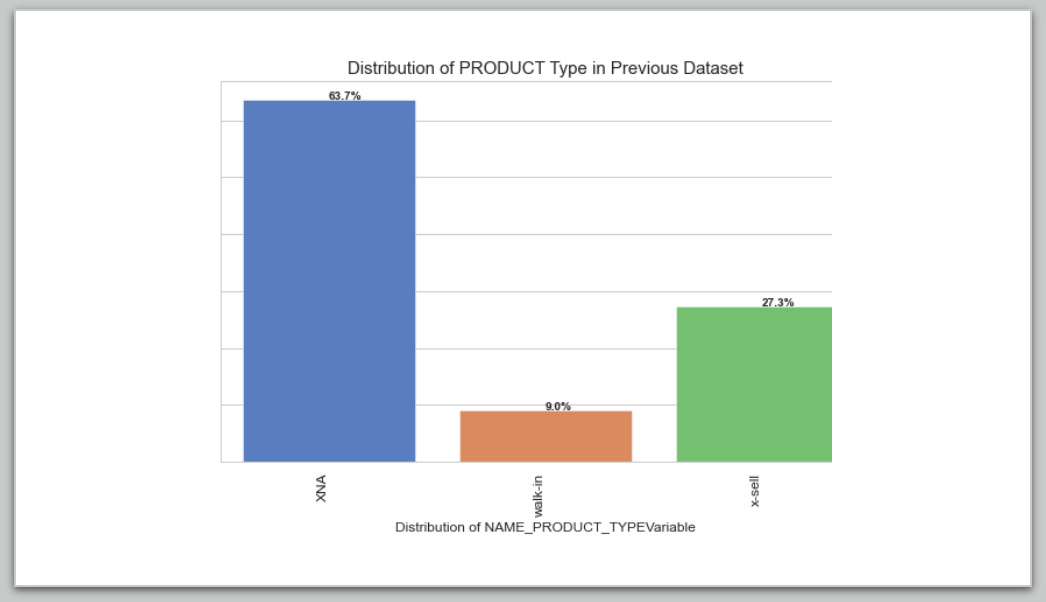
- 41.4% of the total previous applicants have applied for POS & 8.7% is the least applied for cards.
- 41.9% of the Non-Defaulters have applied for POS & the 8.4 % is the least opted for Cards.
- 36.7% of the Defaulters have applied for POS & 10.1% have applied for the Cards.

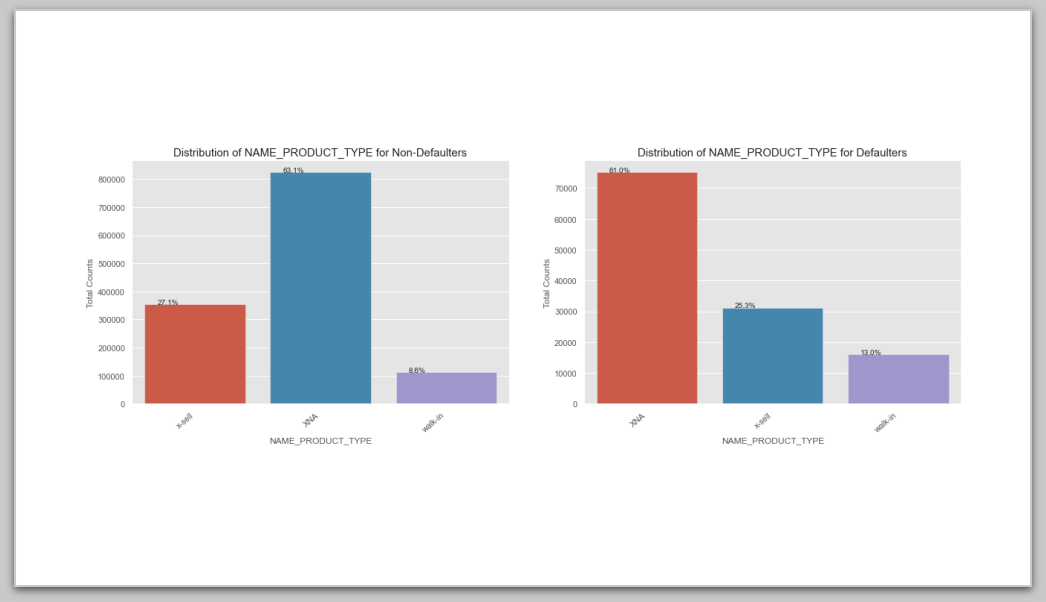




NAME_PRODUCT_TYPE

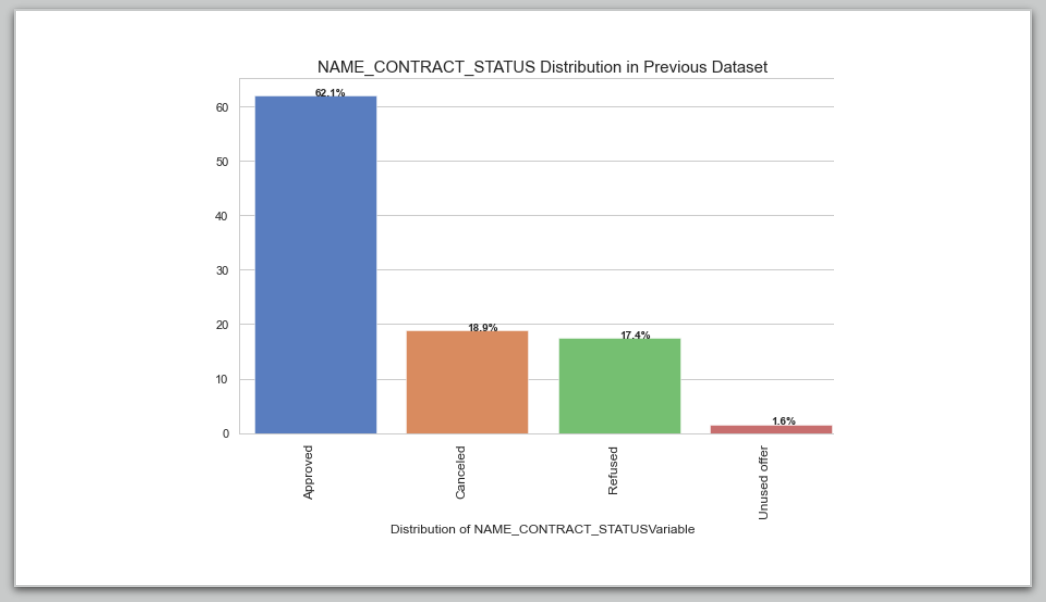
- 63.7% of the Previous applicants have applied for XNA & least is observed for Walk-in of 9.0%.
- 63.1% of the Non-Defaulters have applied for XNA & 8.6% have applied for Walk-in.
- 61.0% of the Defaulters have applied for XNA & 13.0% have applied for walk-in.

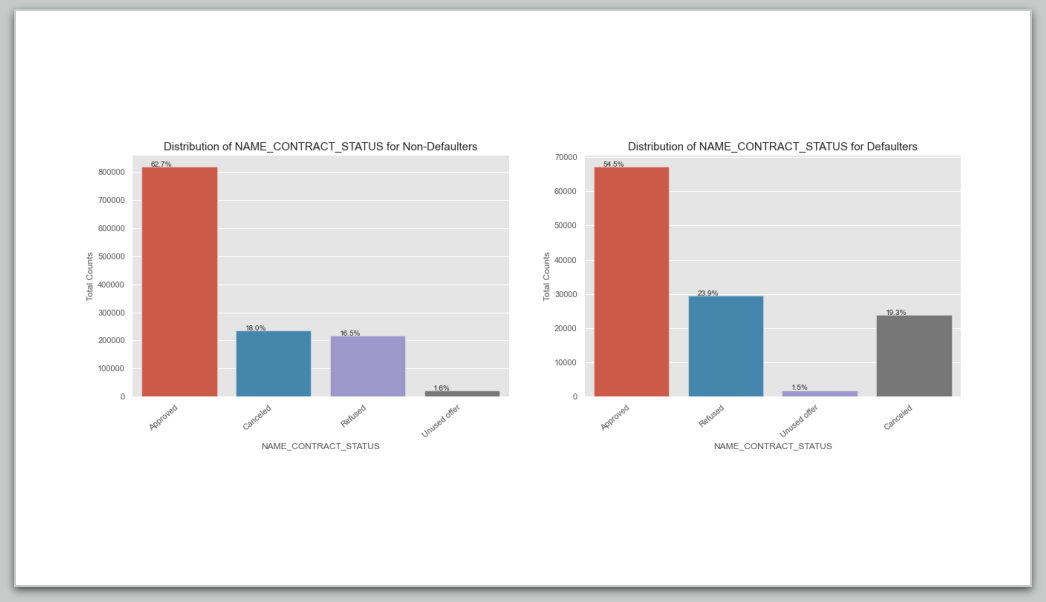




NAME_CONTRACT_STATUS

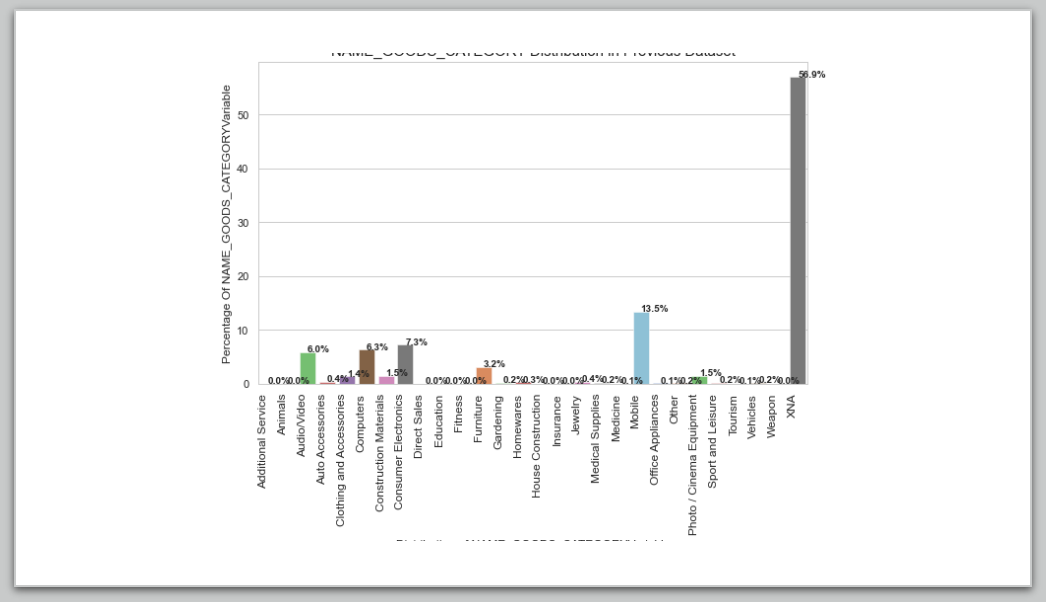
- 62.1% of the total applications have been approved in Previous Dataset.
- 62.7% of the total Non-Defaulters have been approved with Credit in previous credit.
- 54.5% of the total Defaulters have been approved with Credit in the previous credit.

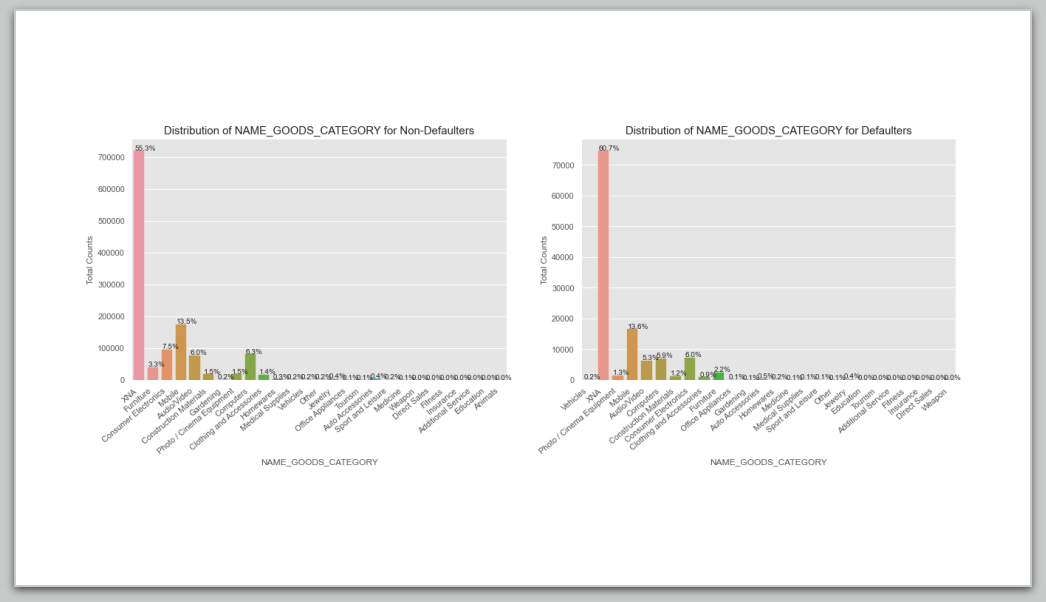




NAME_GOODS_CATEGORY

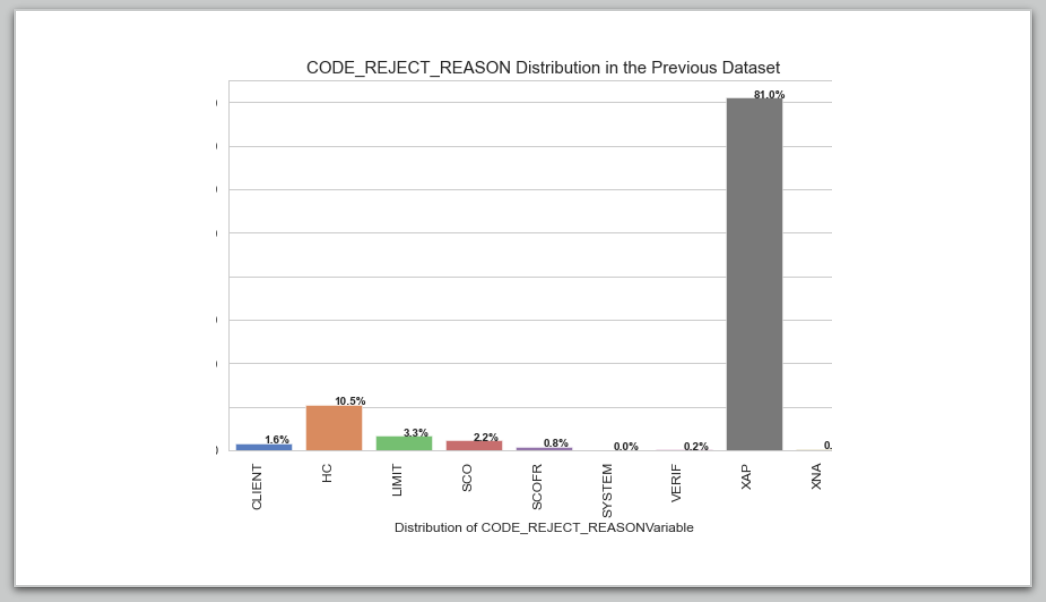
- 56.9% of the total previous applicants have applied for XNA Goods Category.
- 56.3% of the Non-Defaulters have applied for XNA Goods Category.
- 60.7% of the Defaulters also opted for XNA.

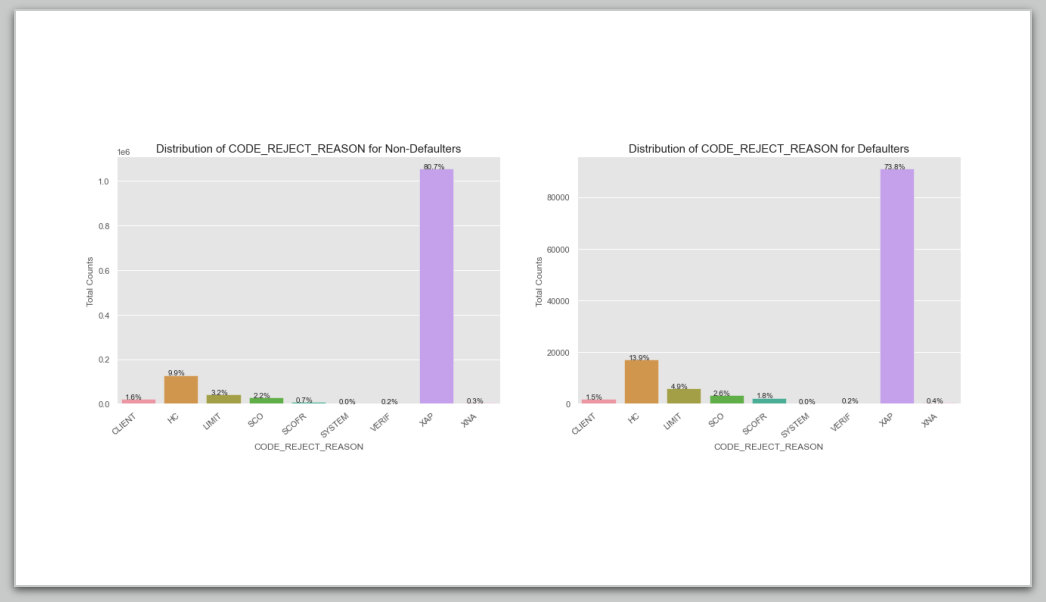




CODE_REJECT_REASON

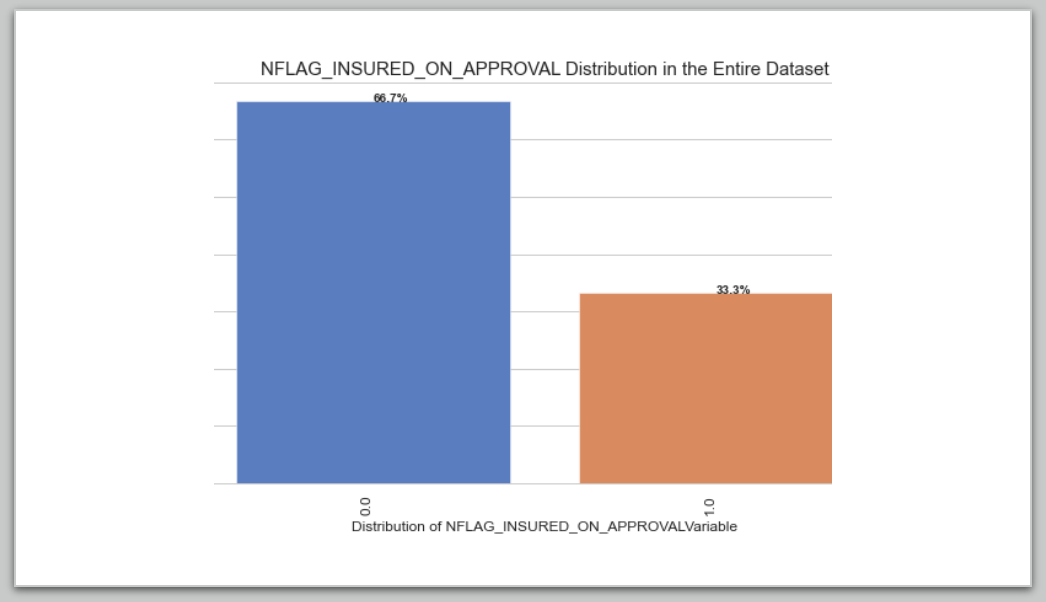
- 81% of the Previous applicants are rejected due to "XAP" and least is 0.2% due to "VERIF".
- 80.7% of the Non-Defaulters are previously rejected due to "XAP" and the least is 0.2% rejected due to "VERIF".
- 73.8% of the Defaulters are previously rejected due to "XAP" and the least 0.2% are rejected due to "VERIF".

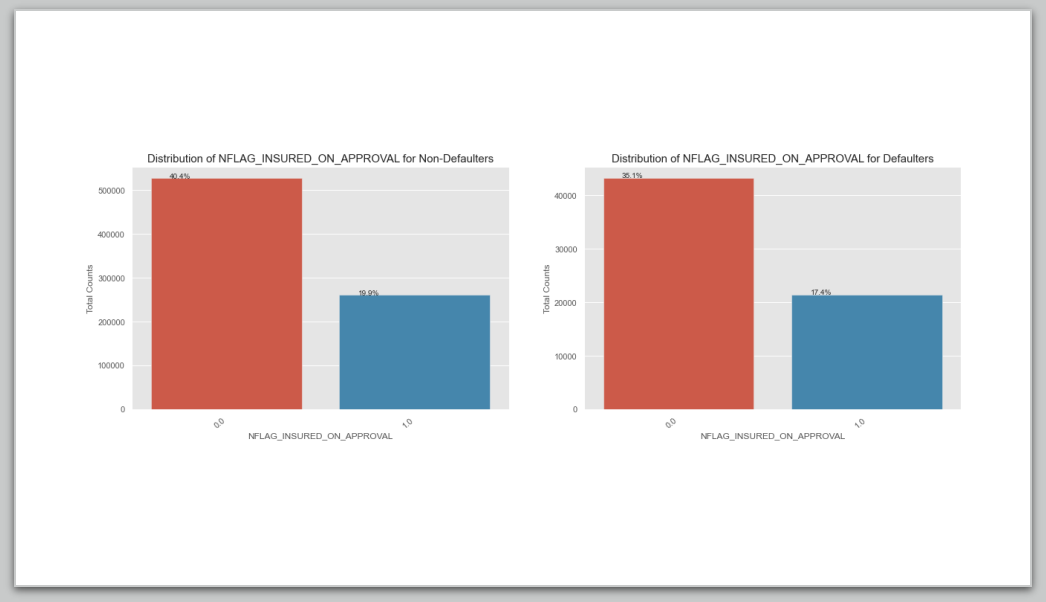




NFLAG INSURED ON APPROVAL

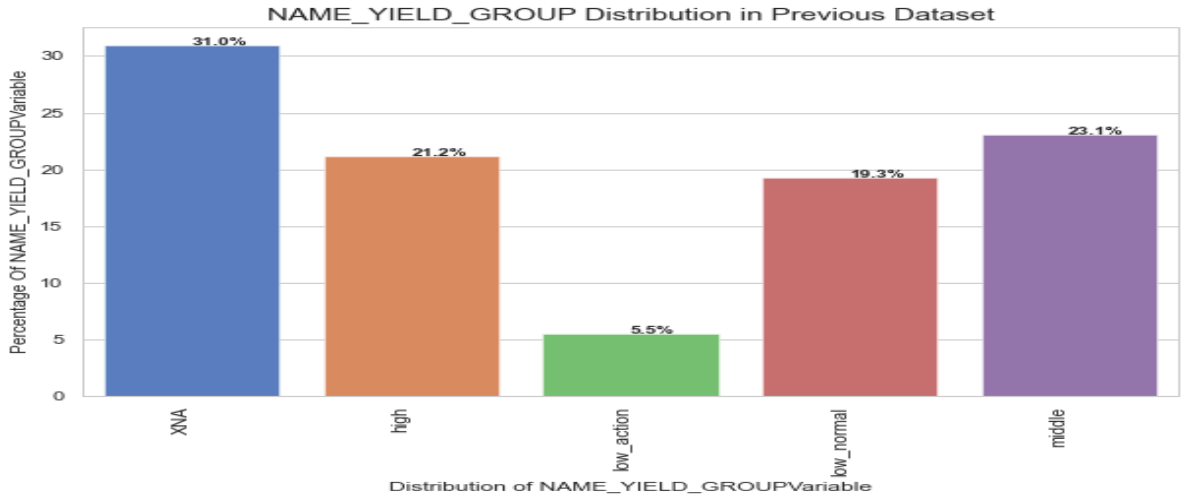
- Only 33.3% of the Previous applicants have applied for Insurance & 66.7% have not applied for Insurance.
- Only 19.9% of the Non-Defaulters previously applied for Insurance & 40.4% have not applied for insurance.
- Only 17.4% of the Defaulters previously applied for Insurance & 35.1% have not applied for insurance.

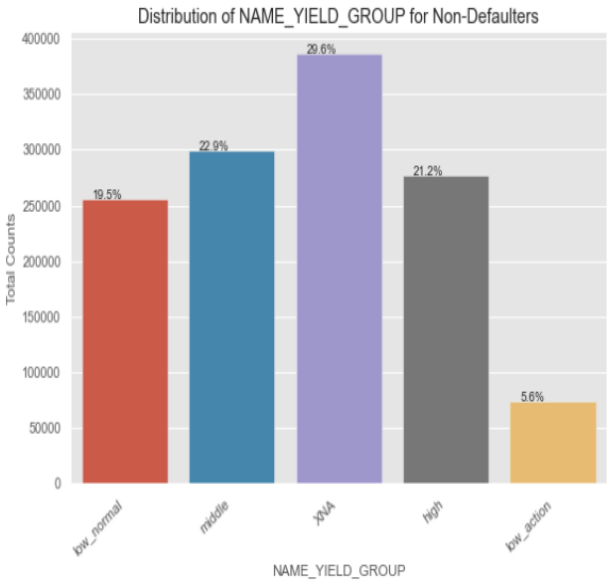


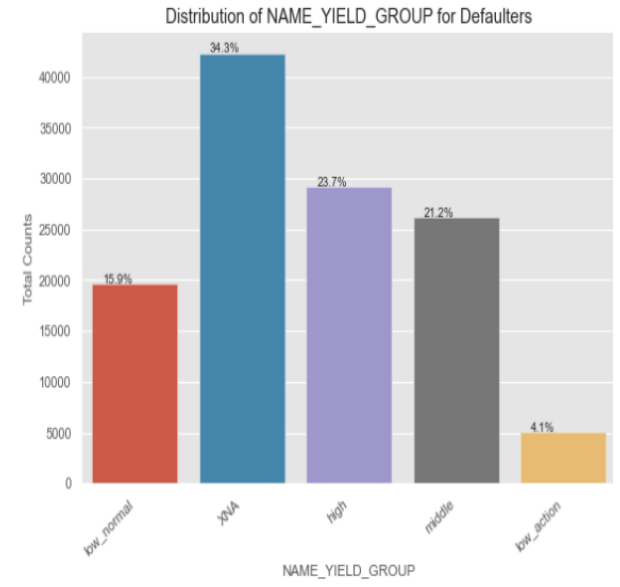


NAME YIELD GROUP

- 1% of the Previous applicants are in the "XNA" interest group. only 5.5% of the total previous applicants are in "bw-action" interest group.
- 29.6% of the Non-Defaulters are in "XNA" group and only 5.6% are in "bw-action" interest group.
- 34.3% of the Defaulters are in "XNA interest group and only 4.1% are in "bw-action" interest group.

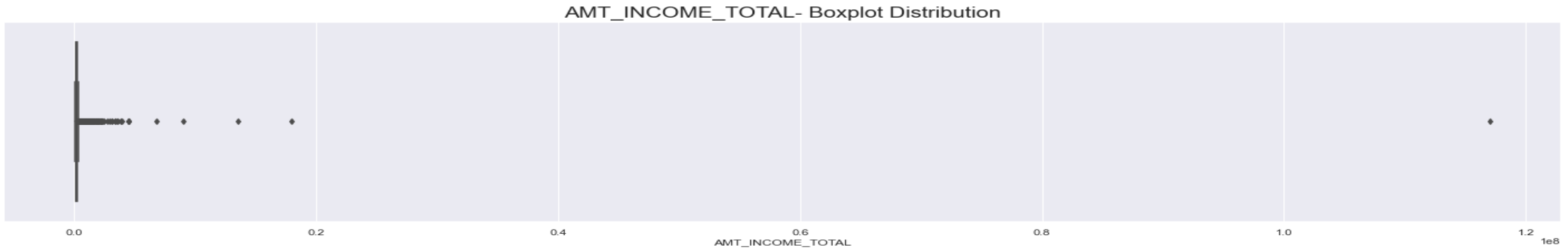






Univariate Analysis Numerical Application Dataset.

Outliers Detection in Amount Income:

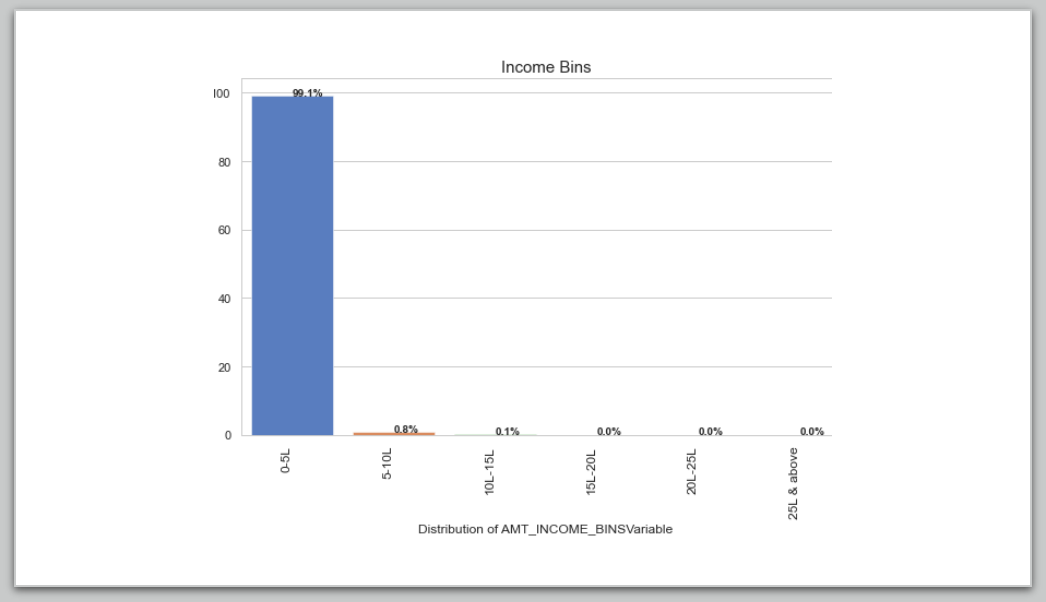


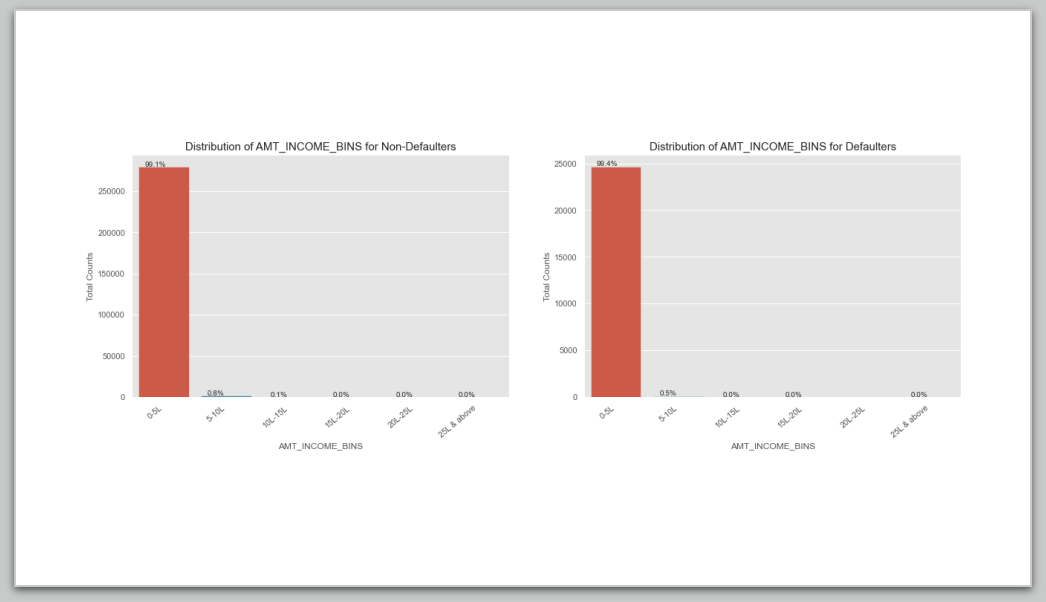
Variable Name	25th Percentile	50th Percentile	Standard Deviation	Mean Value	Median value	75th Percentile	Maximum value
Amount Income Total	112500	147150	237123	168797	147150.0	202500	117000000

- Boxplot summary indicates that most of the values present after the 50th percentile, it means data is right skewed.
- Mean and Median are almost close to each other. This means that it is following normal distribution.
- Datapoints are Clustered at the Upper Whisker and cannot be considered as Outliers as there might be high Income for any Business / Organisations & can be capped at 99 percentile as they might deviate from giving the correct analysis.

Amount Income Bins

- Majority of the Credit applicants appear to be in 0-5Lakhs Income Group with 99.1% & 10-15 Lakhs appear to be least with 0.1% of the overall applicants.
- Majority of the Non-Default Clients appear to be in 0-5Lakhs Income Group with 99.1% & 10-15 Lakhs appear to be least with 0.1%.
- Majority of the Default Clients appear to be in 0-5Lakhs Income Group with 99.4% & 5-10 Lakhs appear to be least with 0.5%.





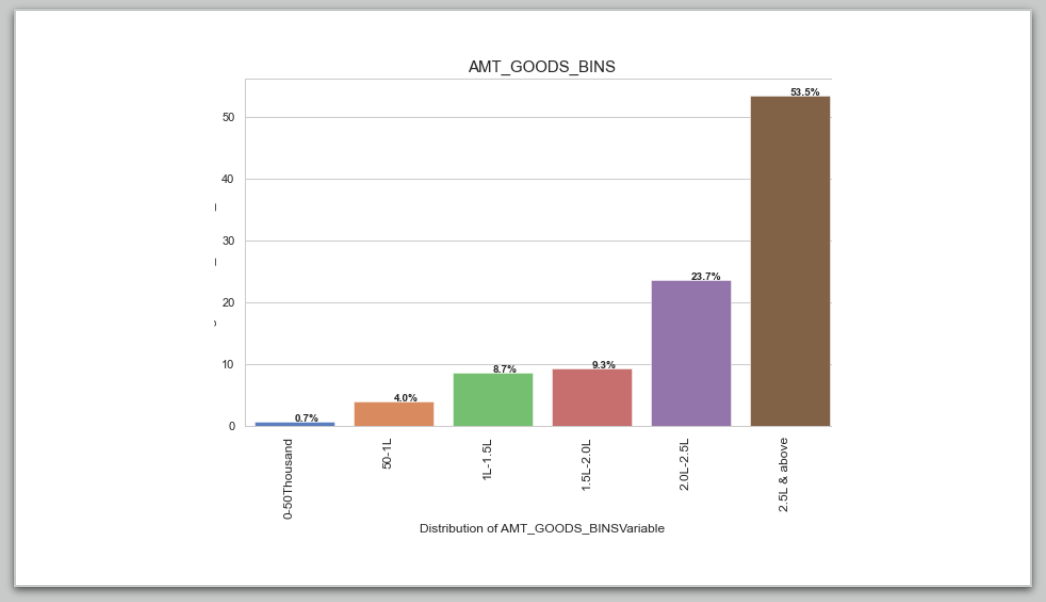
Outliers Detection in Amount Goods Price:

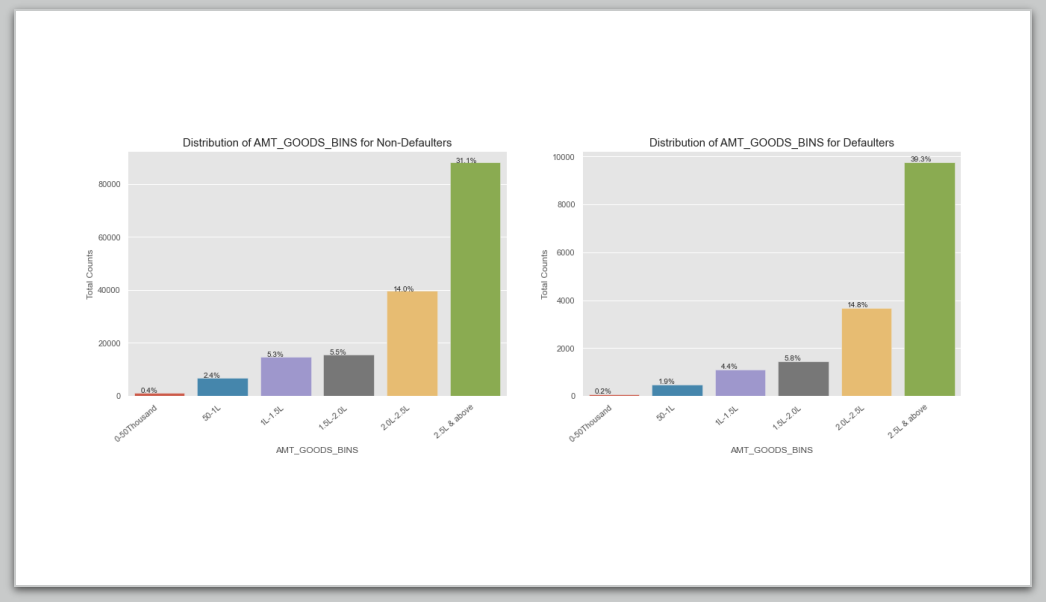
Variable Name	25th Percentile	50th Percentile	Standard Deviation	Mean Value	Median value	75th Percentile	Maximum value
Amount Goods Price	238500	450000	369446	538396	450000	679500	4050000

- Boxplot summary indicates that most of the values present after the 50th percentile, it means data is right skewed.
- Mean and Median are almost close to each other. This means that it is following normal distribution.
- Many Data Points are clustered above Upper Whisker and far away from the upper Whisker
- These points cannot be considered as Outliers as we do not enough source related to the Goods Categories & its Price to prove that these data points are outliers.

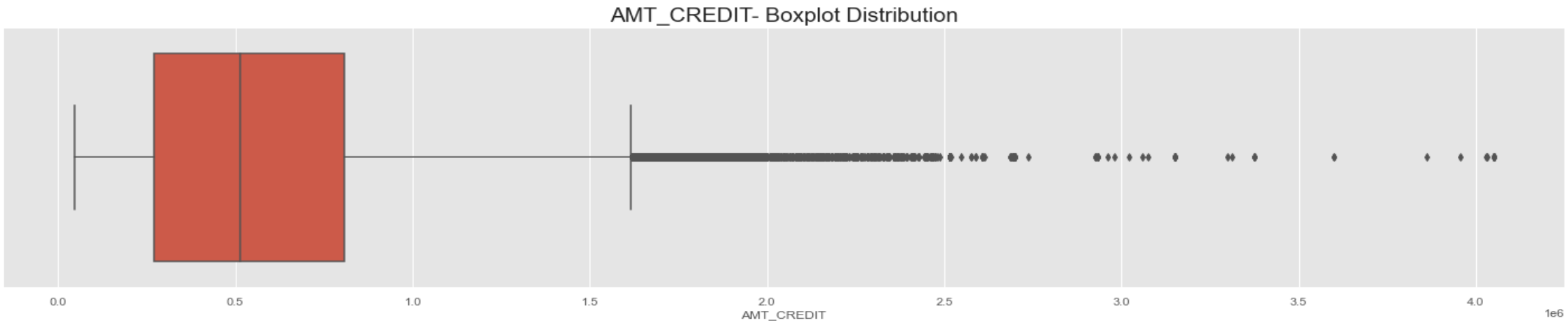
AMT-Goods-Price-Binning

- Majority of the Goods Amount is 2.5Lakhs & above which is 53.5% & least is 0-50 Thousand which is 0.7% of the total applicants.
- Majority of the Non-Defaulters Goods Amount is 2.5Lakhs & above which is 31.1% & least is 0-50 Thousand which is 0.4%.
- Majority of the Defaulters Goods Amount is 2.5Lakhs & above which is 39.3% & least is 0-50 Thousand which is 0.2%.





Outliers Detection in Amount Credit:

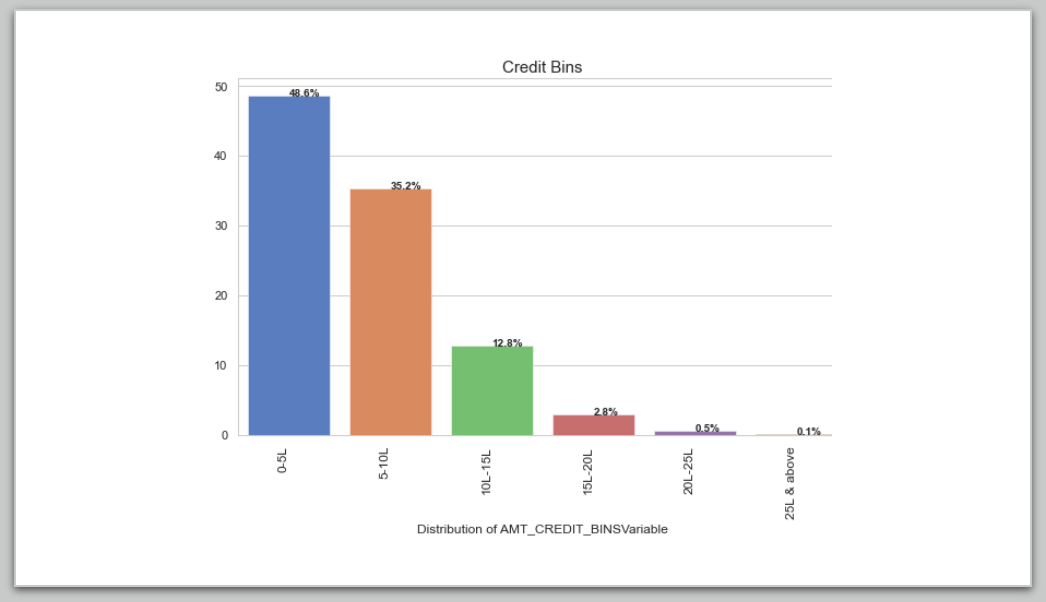


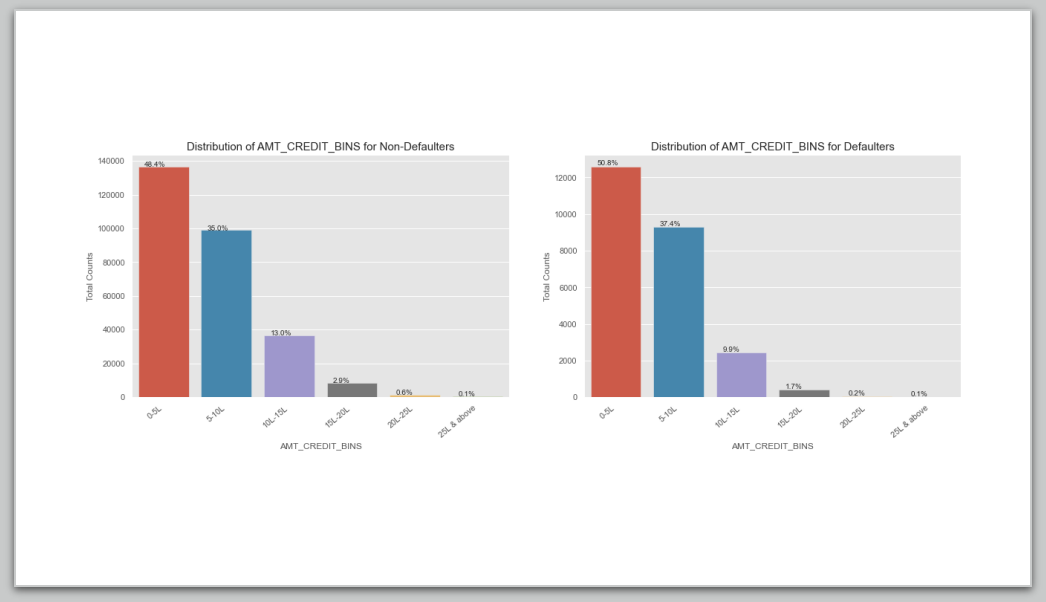
Variable Name	25th Percentile	50th Percentile	Standard Deviation	Mean Value	Median value	75th Percentile	Maximum value
Amount Goods Price	270000	513531	402490	599025	513531	808650	4050000

- Mean and Median are almost far from each other. This means that it is not following normal distribution.
- Many Datapoints are clustered above the Upper Whisker & cannot be considered as Outliers as they might be correct in real times.

Amount Credit-Binning

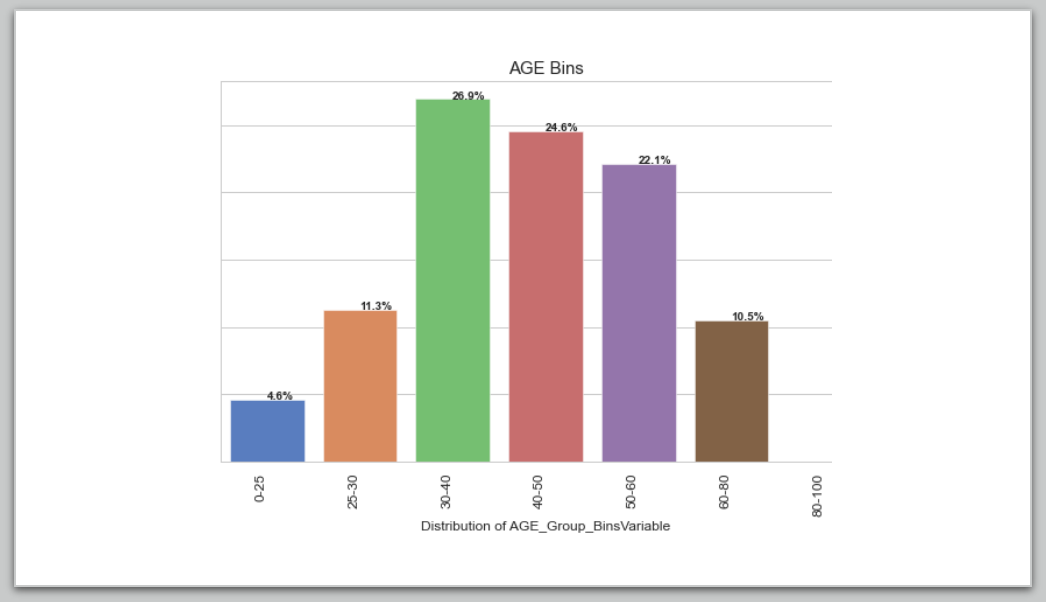
- Majority of the Credit applicants Amount appears to be in 0-5 Lakhs which constitutes to 48.6% and least Credit amount appears to be in 25Lakhs & above with 0.1% of the total applicants.
- Majority of the Non-Defaulters Credit amount is in 0-5 Lakhs range with 48.4% & least Credit Amount is 25Lakhs & above with 0.1%.
- Majority of the Defaulters Credit amount is in 0-5 Lakhs range with 50.8% & least Credit Amount is 25Lakhs & above with 0.1%.

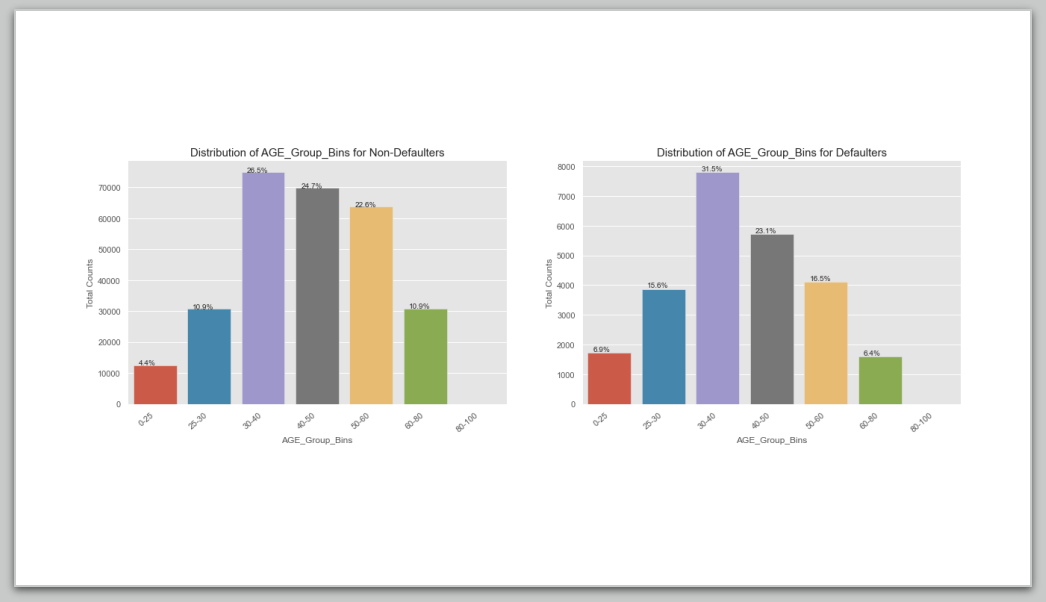




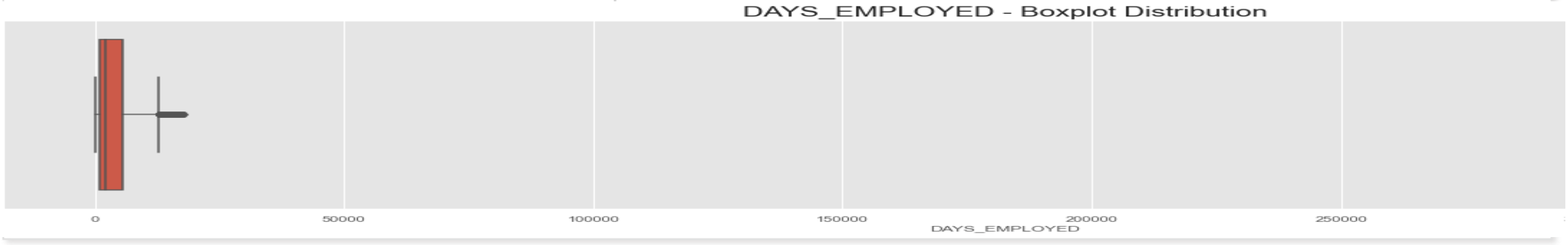
Age Group Binning

- 30-40 age group holds the highest 26.9% of the total applicants & the least is 0-25 age group with 4.6%.
- 26.5% of the Non-Defaulters is 30-40 age group and the least is 4.4% of the 0-25 age group.
- 31.5% of the Defaulters is 30-40 age group & the least is 60-80 age group with 6.4% of the total defaulters.





Outliers in Days Employed



Variable Name	25th Percentile	50th Percentile	Mean Value	Median value	75th Percentile	Maximum value
Days Employed	933 days	2219 days	67724 days	2219 days	5707 days	365243 days

- Some data points are clusters above the Upper Whisker & one data point is very far from the Upper Whisker.
- Max Value of Days Employed is 365243.0 insignificant as Days Employed cannot exceed the 1000 years.
- Capping is the best technique to treat these Outliers to replaced with 99 percentile value.

Correlation Analysis

Top 10 Correlation Scores - Target -0 - Non- Defaulters- Application Dataset

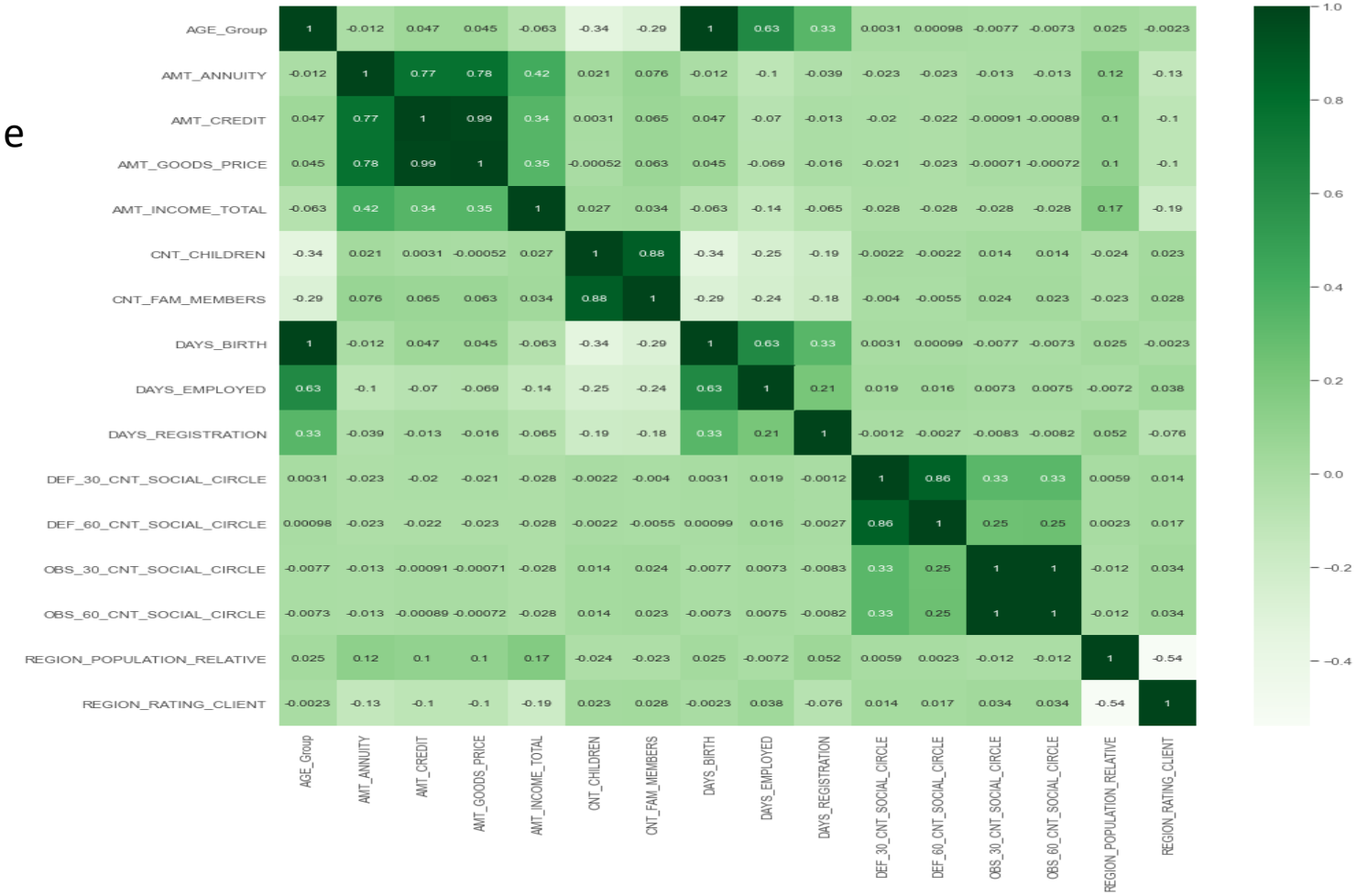
Variable Name-1	Variable Name-2	Pearson Correlation Value
AMT_GOODS_PRICE	AMT_CREDIT	0.99
AMT_GOODS_PRICE	AMT_ANNUITY	0.77
CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
AMT_ANNUITY	AMT_CREDIT	0.77
DAYS_EMPLOYED	DAYS_BIRTH	0.63
DAYS_REGISTRATION	DAYS_BIRTH	0.33
AMT_ANNUITY	AMT_INCOME_TOTAL	0.42

Top 10 Correlation Values - Target - 0 - Non Defaulters- Application Dataset

From the Non Defaulters Correlation map,

the Variables that be considered to have the highest Pearson Correlation are:

1. AMT_CREDIT
2. AMT_GOODS_ PRICE
3. AMT_INCOME_TOTAL
4. AMT_ANNUITY
5. DAYS_EMPLOYED
6. DAYS_BIRTH
7. DAYS_REGISTRATION
8. CNT_FAM_MEMBERS
9. CNT_CHILDREN



Top 10 Correlation Scores - Target-1 – Defaulters Application Dataset

Variable Name-1	Variable Name-2	Pearson Correlation Score
AMT_GOODS_PRICE	AMT_CREDIT	0.98
AMT_GOODS_PRICE	AMT_ANNUITY	0.75
CNT_FAM_MEMBERS	CNT_CHILDREN	0.89
AMT_ANNUITY	AMT_CREDIT	0.75
REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	0.44
DAYS_EMPLOYED	DAYS_BIRTH	0.58

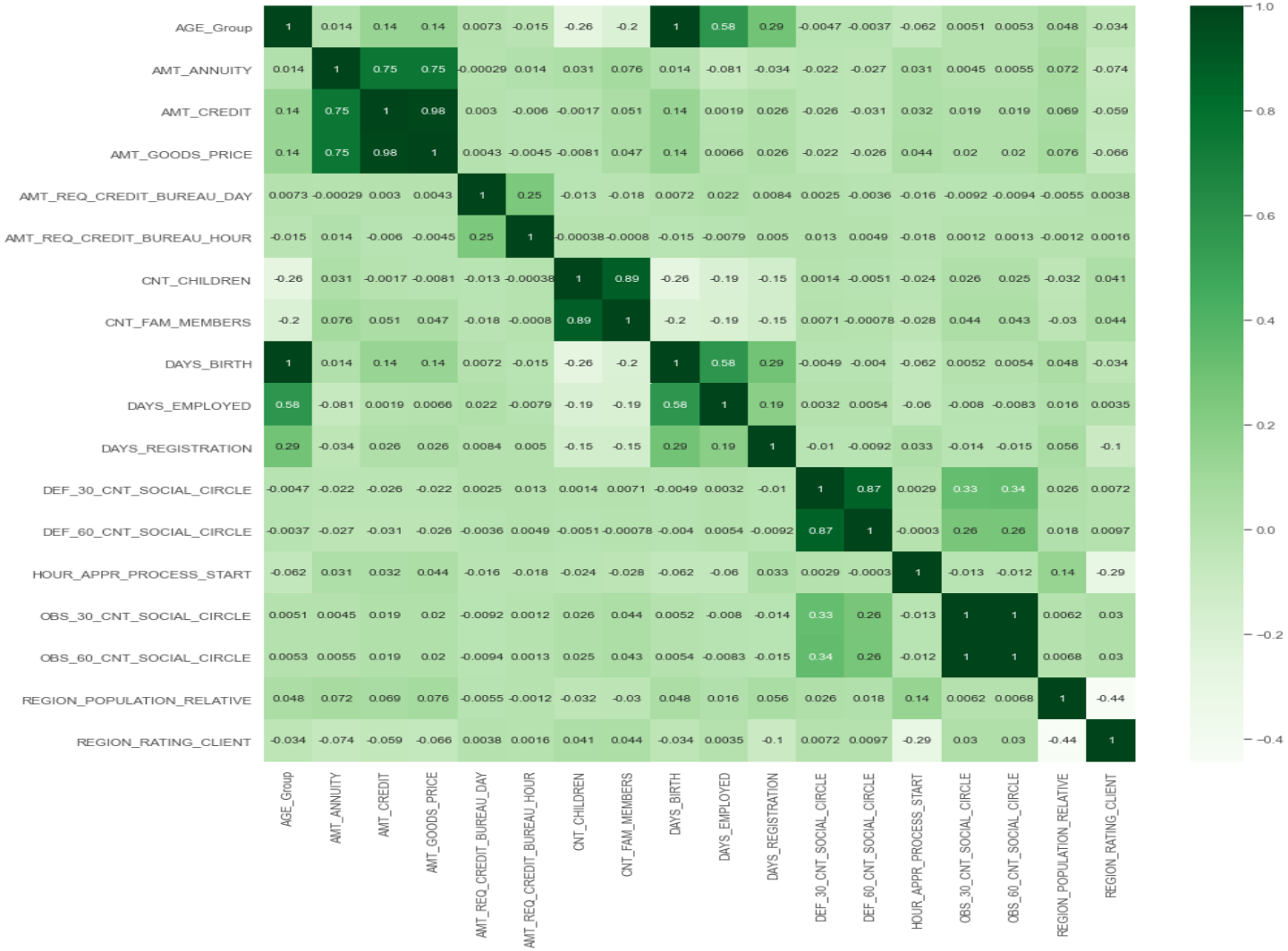
Top 10 Correlation Values - Target-1 – Defaulters Application Dataset

From the above Defaulters Correlation

Heatmap, the top Variables we can consider

as per the highest Pearson Correlation are :

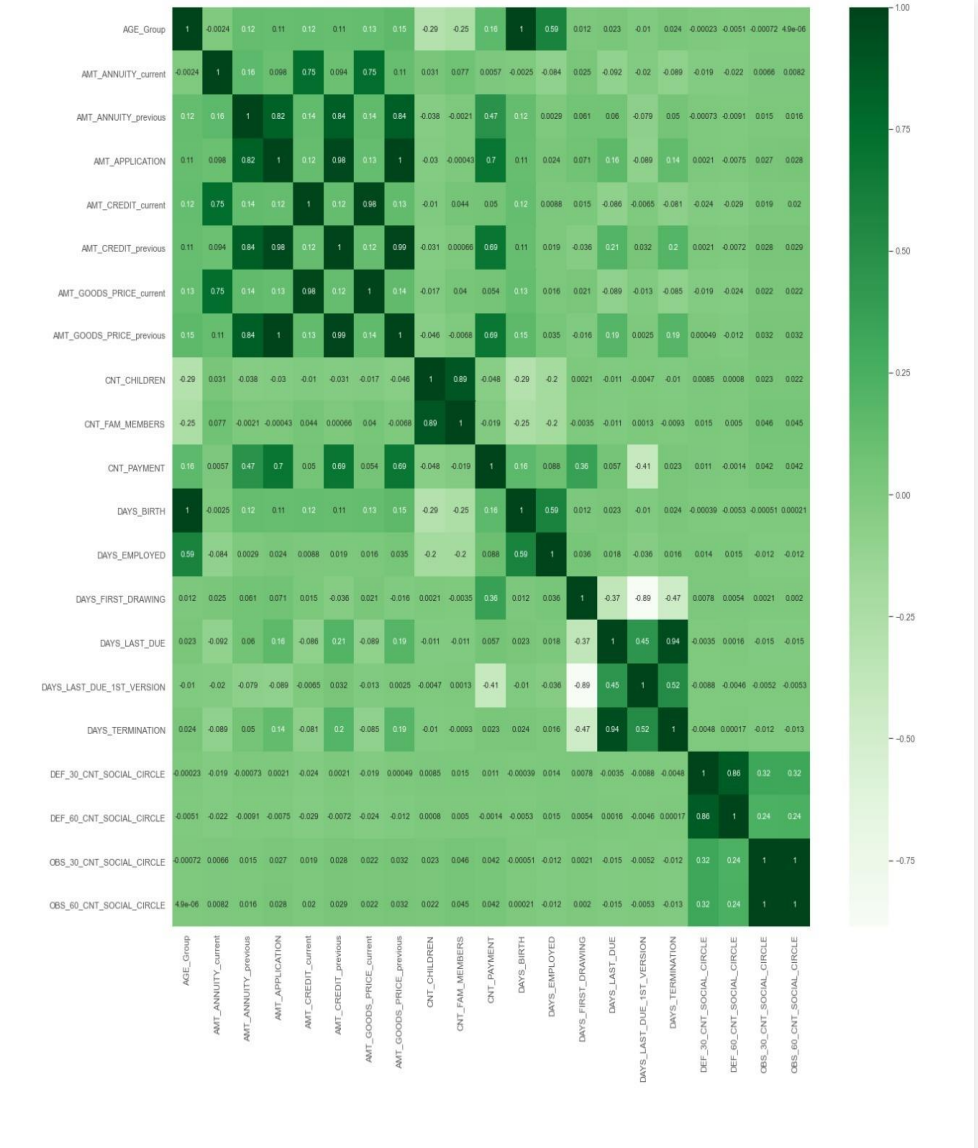
- AMT_CREDIT
- AMT_GOODS_ PRICE
- CNT_FAM_MEMBERS
- CNT_CHILDREN
- AMT_ANNUITY
- DAYS_EMPLOYED
- DAYS_BIRTH
- REGION_RATING_CLIENT
- REGION_POPULATION_RELATIVE



Top 10 Correlation Values - Target - 0 - Non Defaulters- Merged Dataset

- The top 10 Correlated variables for Non-Defaulters is::

1. AMT_APPLICATION – AMT-CREDIT-previous
2. AMT-GOODS-PRICE-previous – AMT-CREDIT-previous
3. DAYS_LAST_DUE - DAYS_TERMINATION
4. AMT-CREDIT-current –AMT-GOODS-PRICE-current
5. CNT_CHILDREN - CHT_FAMILY_MEMBERS
6. AMT-ANNUITY-previous – AMT-GOODS-PRICE-previous
7. AMT-ANNUITY-previous – AMT-CREDIT-previous
8. AMT-ANNUITY-current – AMT-CREDIT-current
9. CNT_PAYMENT - AMT_APPLICATION
10. CNT_PAYMENT – AMT GOODS-PRICE-previous



Top 10 Correlation Values - Target - 1 - Defaulters- Merged Dataset

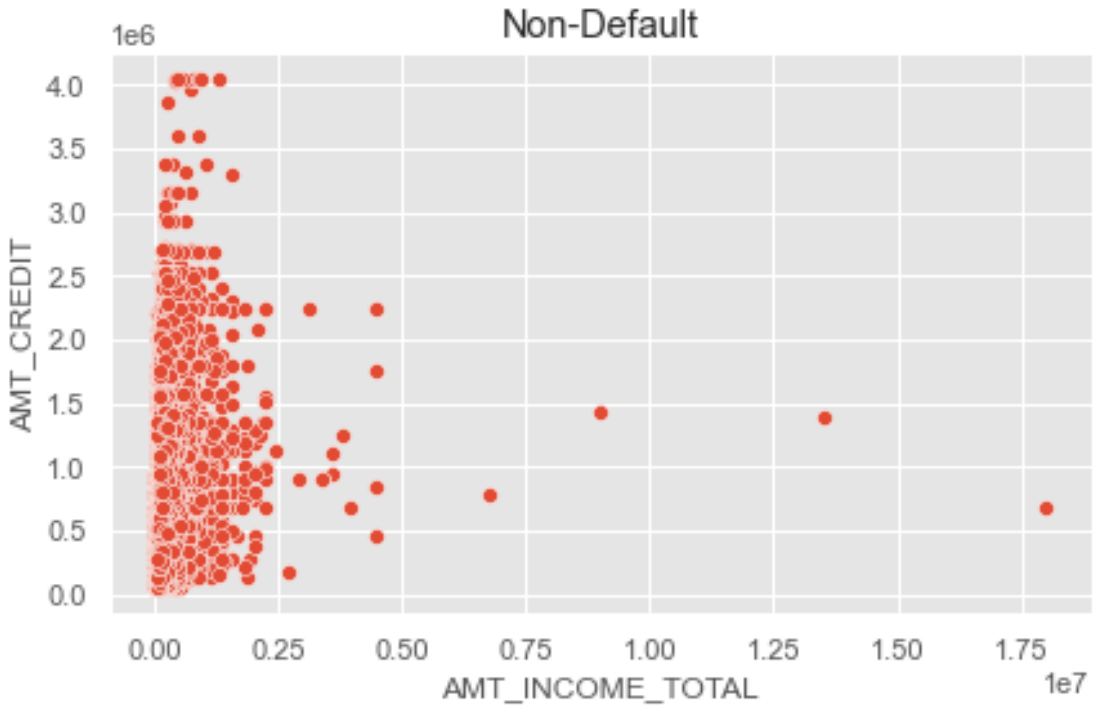
The top 10 Correlated variables for Defaulters is :

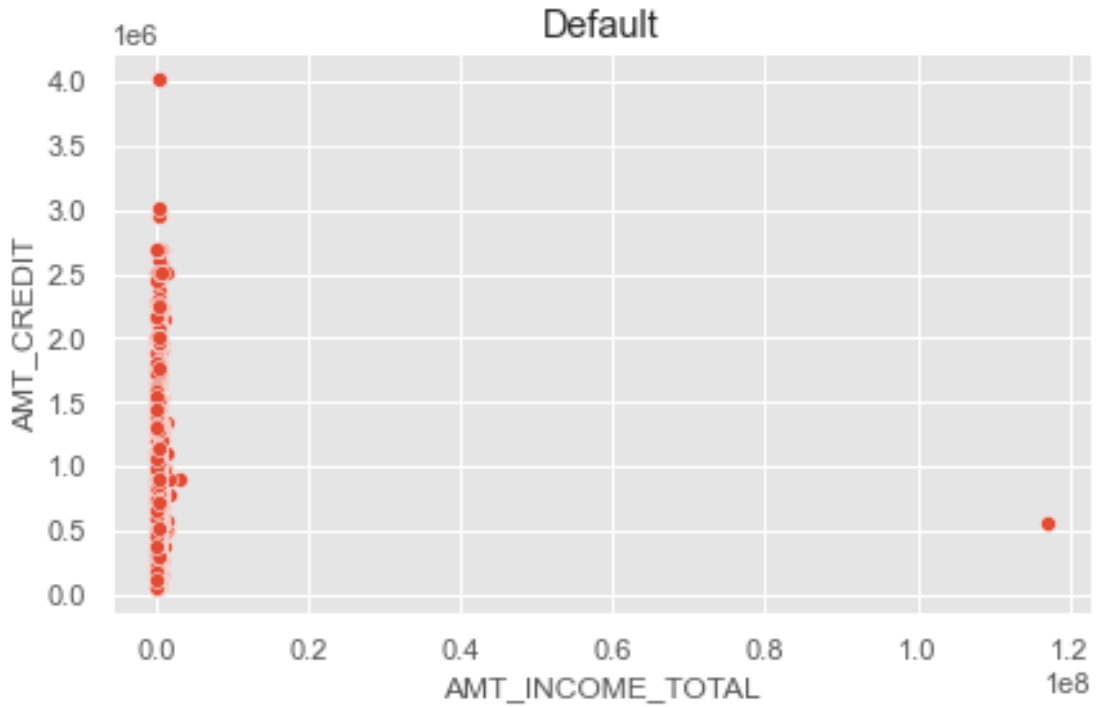
1. AMT APPLICATION – AMT CREDIT previous
2. AMT GOODS PRICE previous – AMT CREDIT previous
3. DAYS_LAST_DUE - DAYS_TERMINATION
4. AMT CREDIT_ current – AMT GOODS PRICE current
5. CNT_CHILDREN - CHT_FAMILY_MEMBERS
6. AMT ANNUITY-previous – AMT GOODS-PRICE-previous
7. AMT ANNUITY-previous – AMT CREDIT-previous
8. AMT ANUUITY-current – AMT CREDIT-current

Bivariate Analysis Numerical Application Dataset.

Amount Credit vs Amount Income

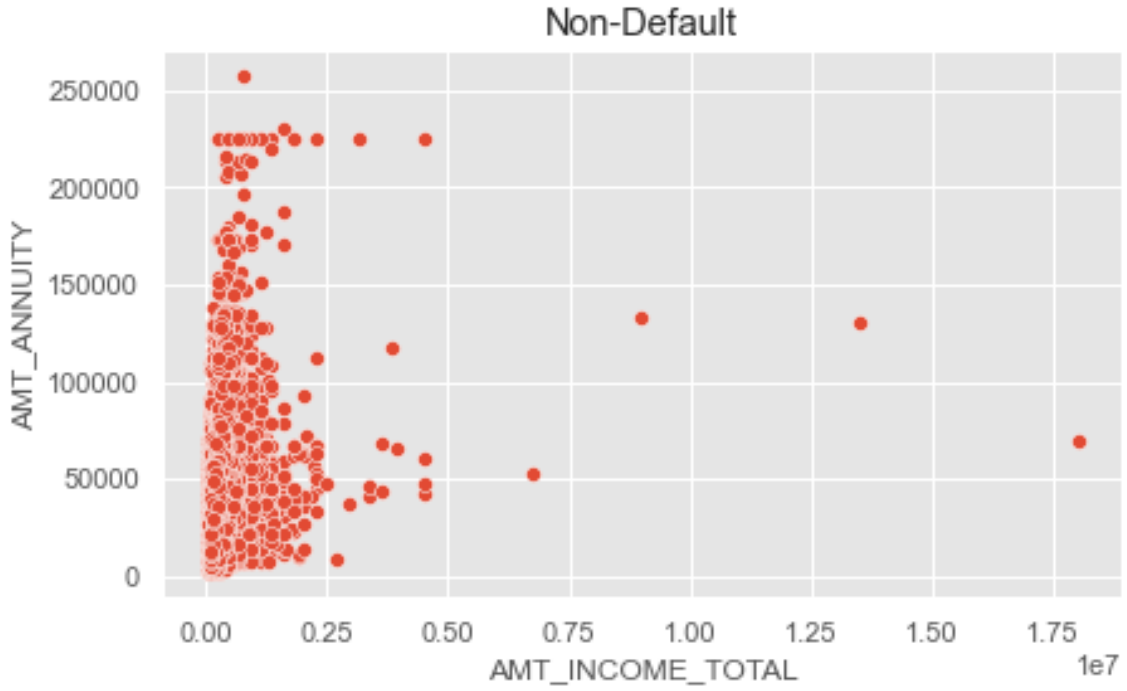
- Among Non-Defaulters, Income & Credit amounts are clustered and only few data points are scattered &have positive linear relationship.
- Among Defaulters, Income & Credit is Clustered and no linear relationship observed.

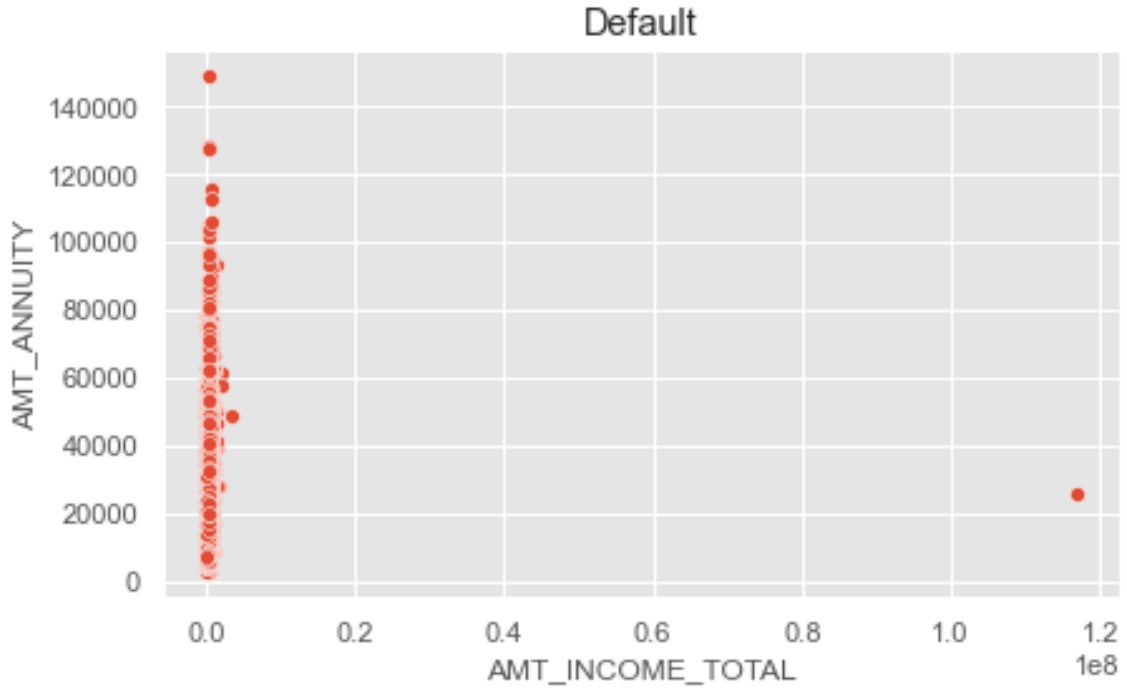




Amount Income Total vs Amount Annuity

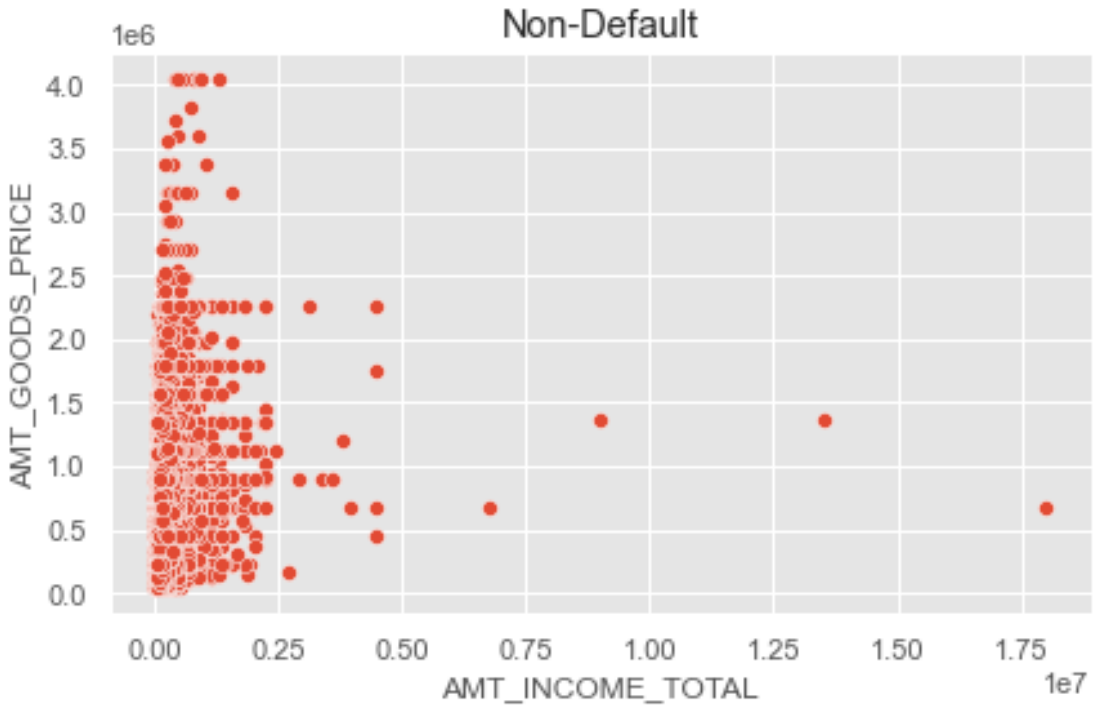
- Among Non-Defaulters, Income & Annuity amounts are clustered and only few data points are scattered & have positive linear relationship.
- Among Defaulters, Income & Annuity is Clustered & no linear relationship observed.

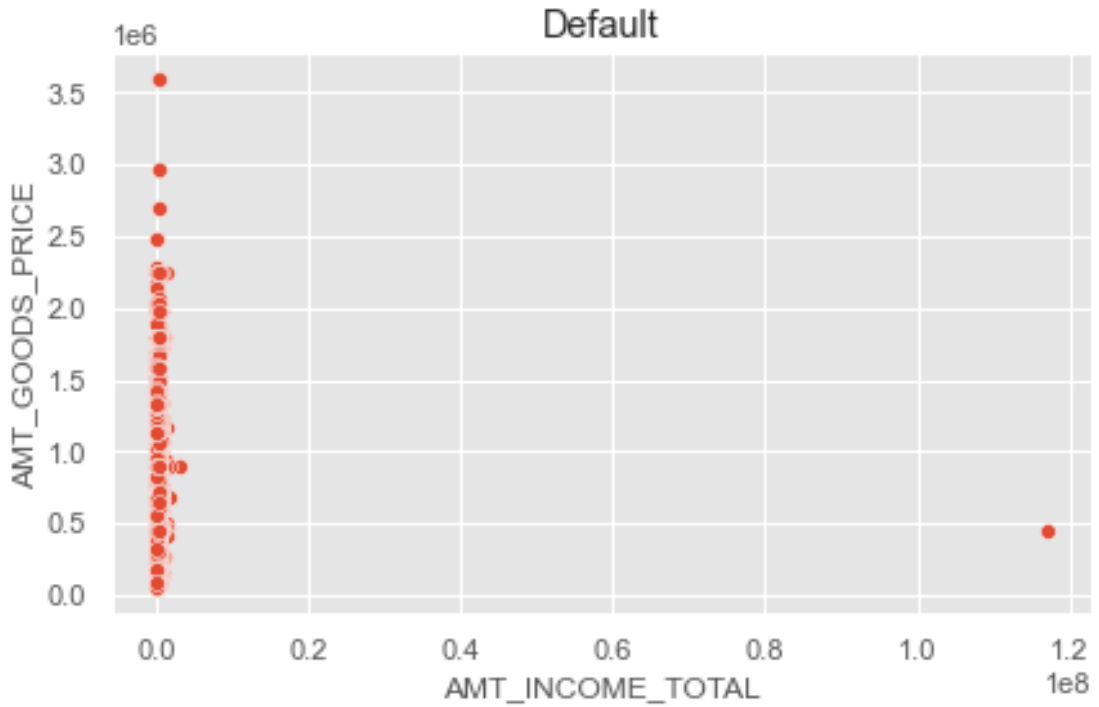




Amount Income Total vs Amount Goods Price

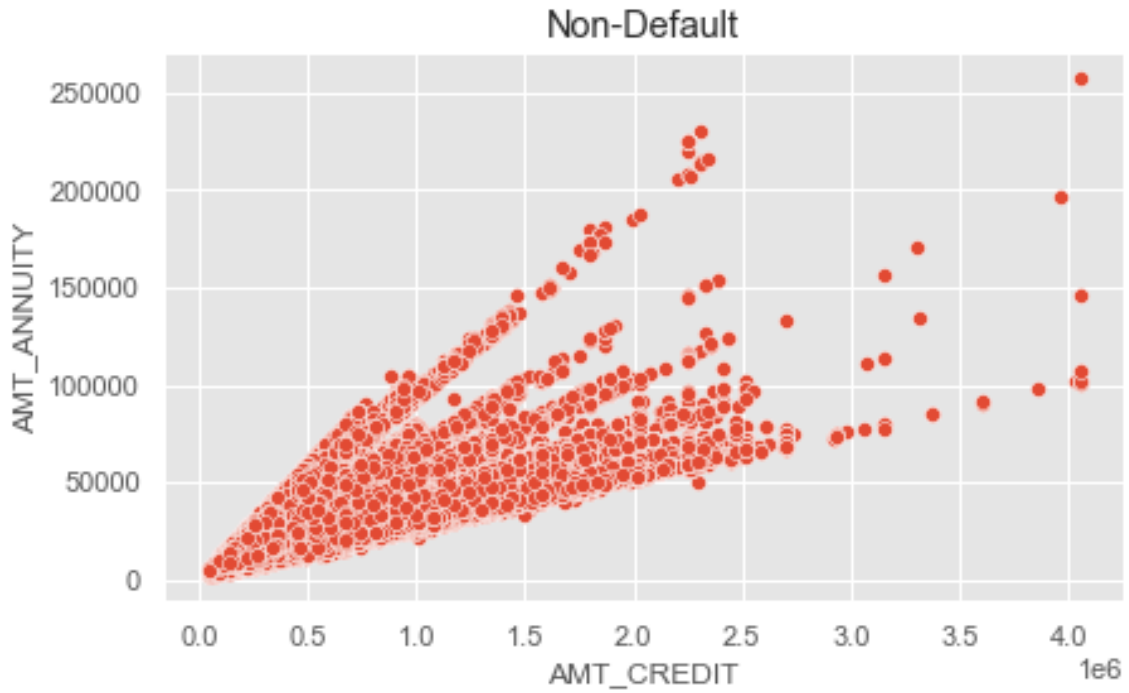
- Among Non-Defaulters, Income & Goods Price amounts are clustered and only few data points are scattered & have positive linear relationship.
- Among Defaulters, Income & Goods Price amounts is Clustered & no linear relationship observed.

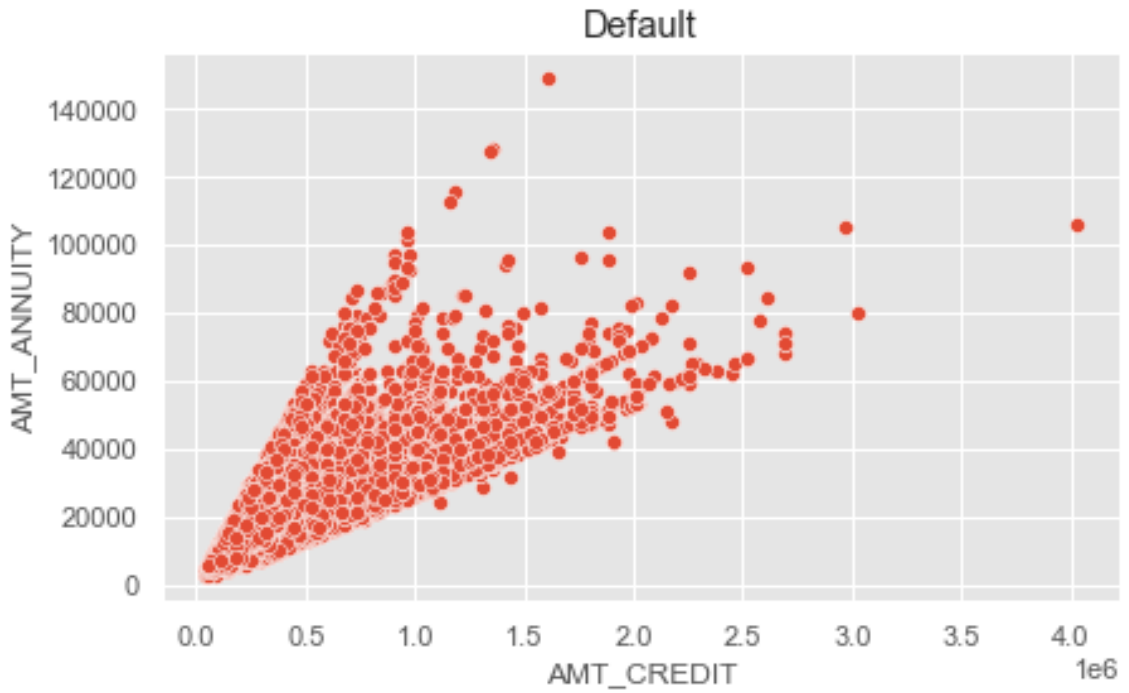




Amount Credit vs Amount Annuity

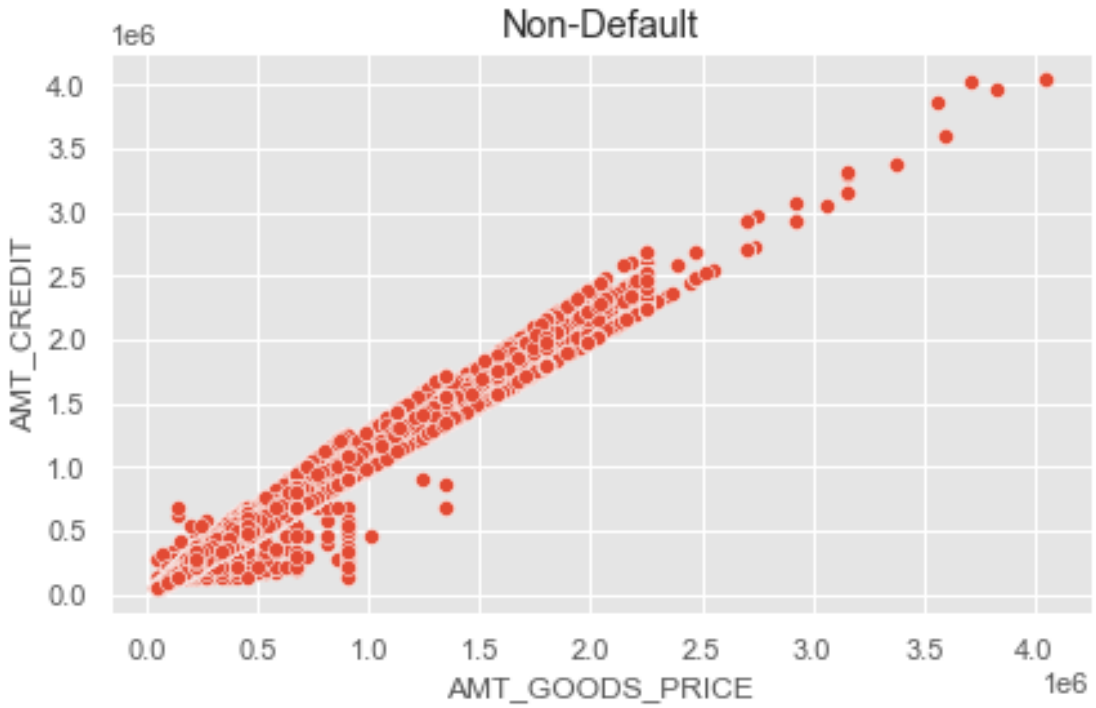
- Among Non-Defaulters, Credit Amount & Annuity show a positive linear relationship, as the amt Credit is increasing, there is a positive increase in the Annuity.
- Among Defaulters, Credit Amount & Annuity show a positive linear relationship.
- As the Credit amount increases, Annuity amount also increases, few points are little scattered.

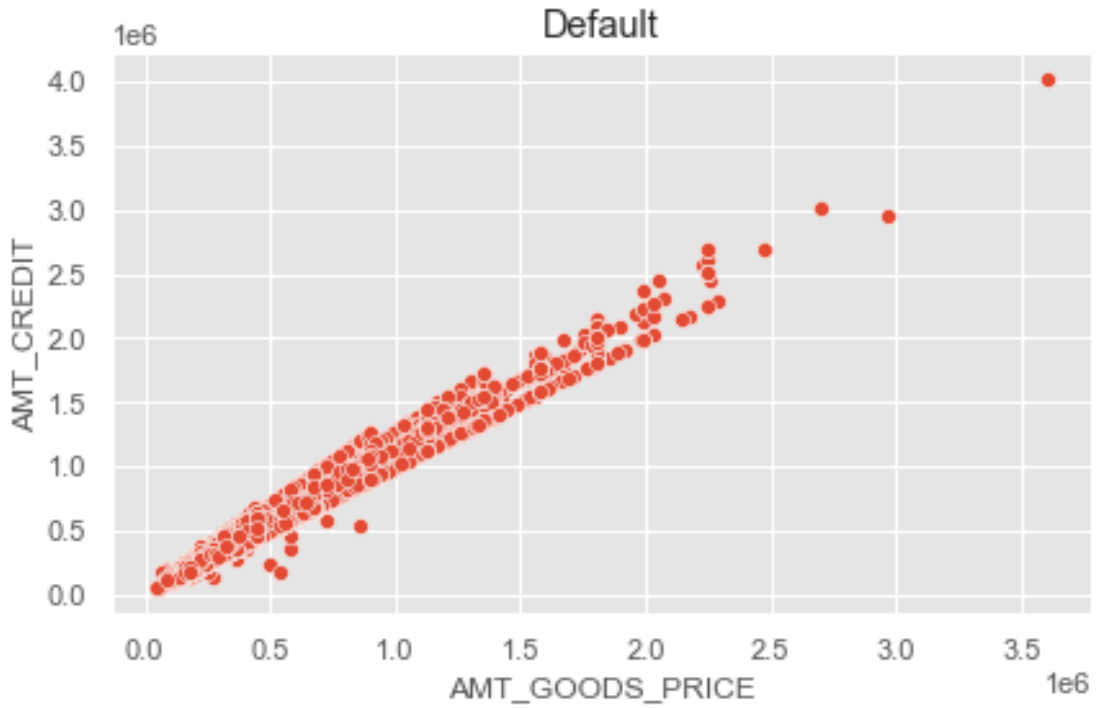




Amount Goods Price vs Amount credit

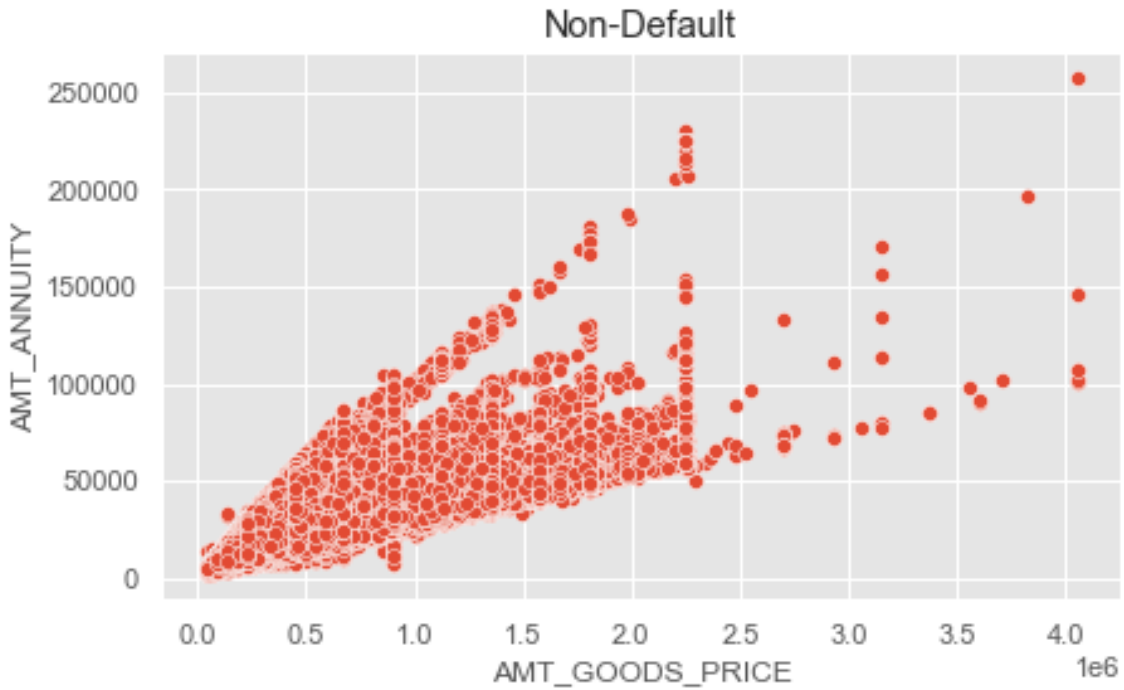
- For Non-Defaulters, Amount Goods Price & Credit show positive linear relationship.
- Defaulters also have a positive linear relationship between Credit & Goods Price.





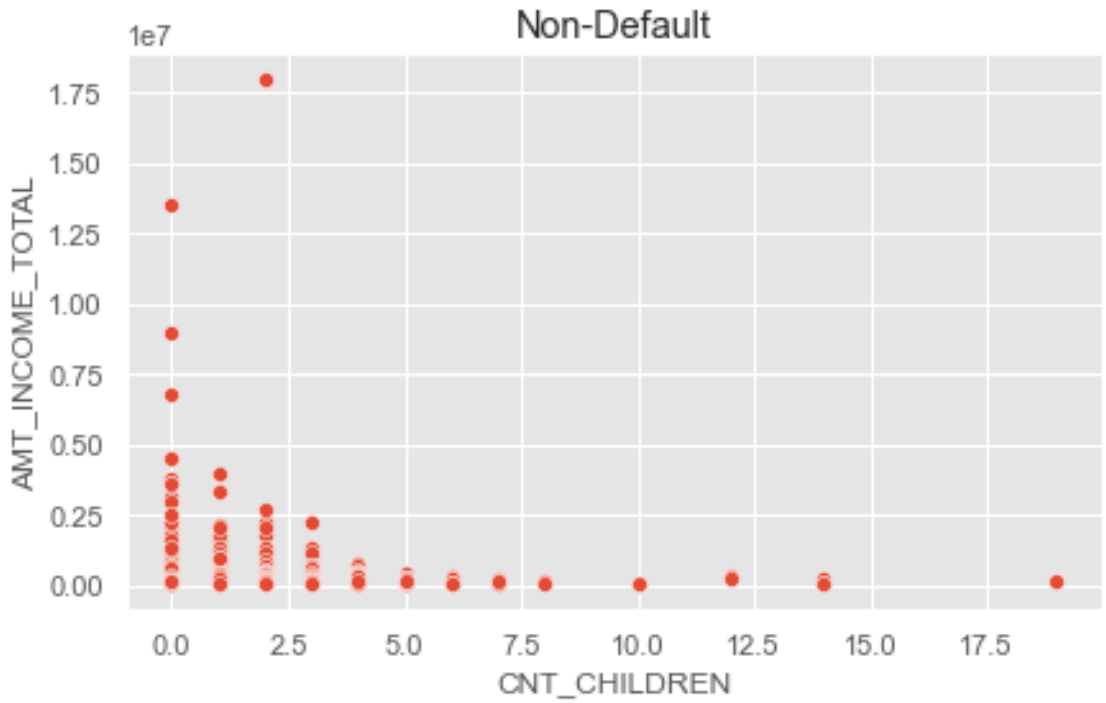
Goods Price vs Annuity

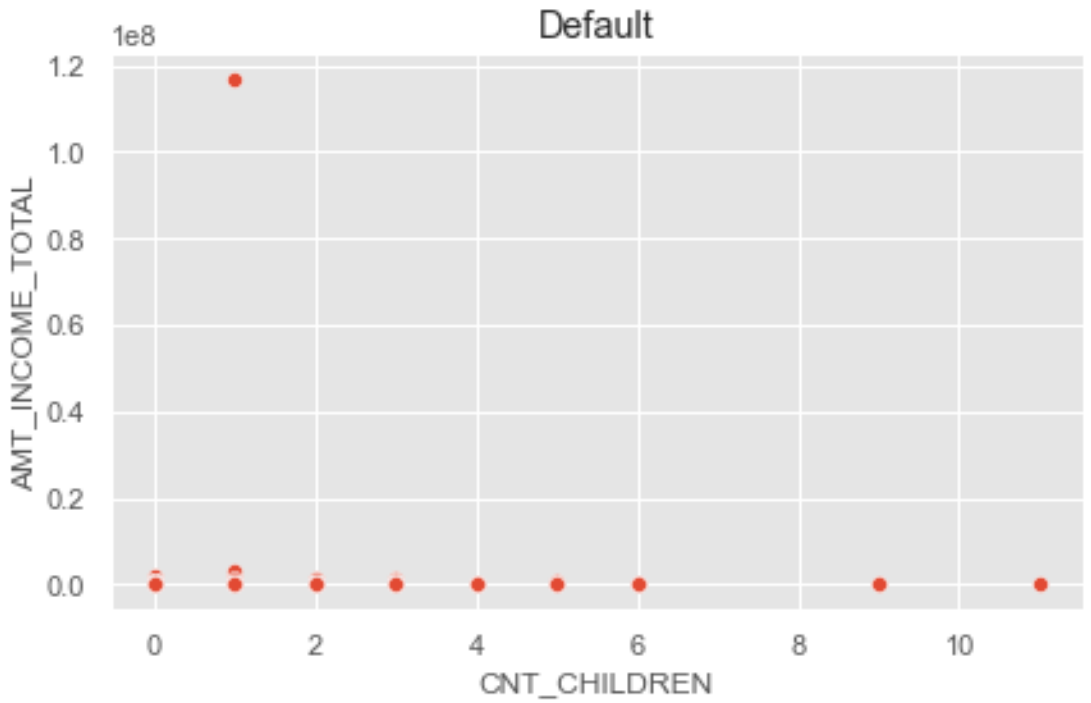
- Among Non-Defaulters, the Goods Price & Annuity show Positive linear relationship.
- Defaulters also show Positive linear relationship between Goods Price & Annuity.



Count of children vs Amount Income Total

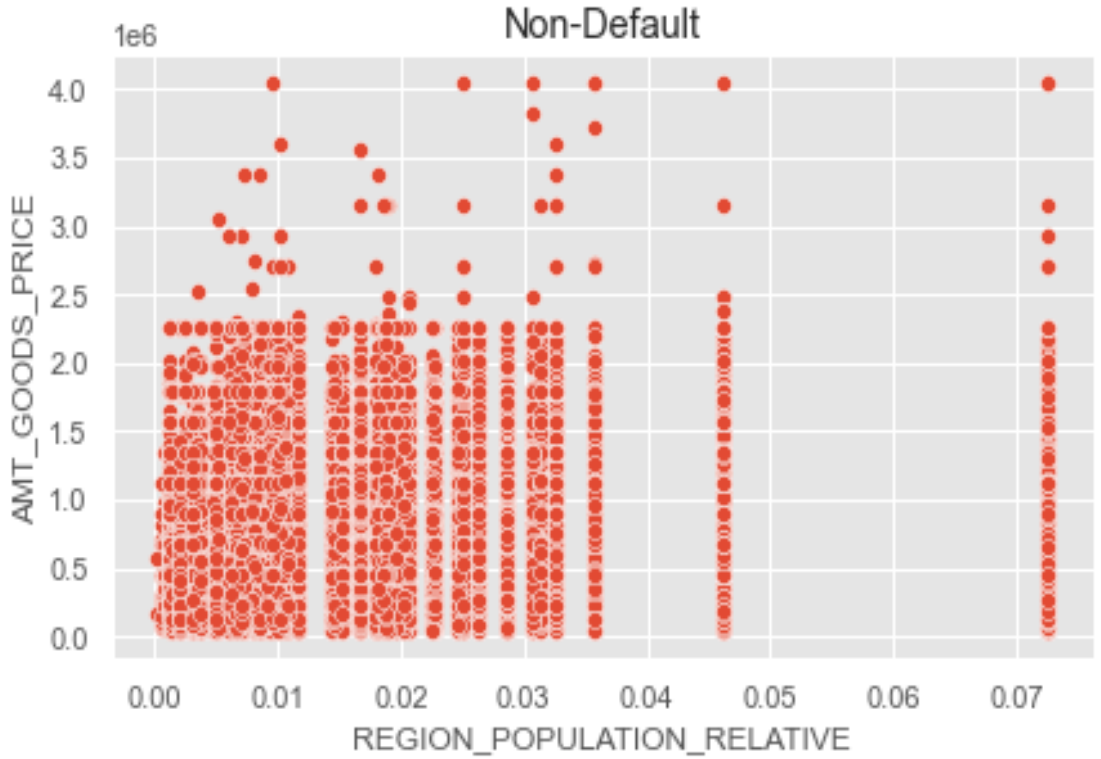
- Defaulters have less children & less Income around when compared to Non-Defaulters.
- Also there is no positive relationship between the children count & Income.

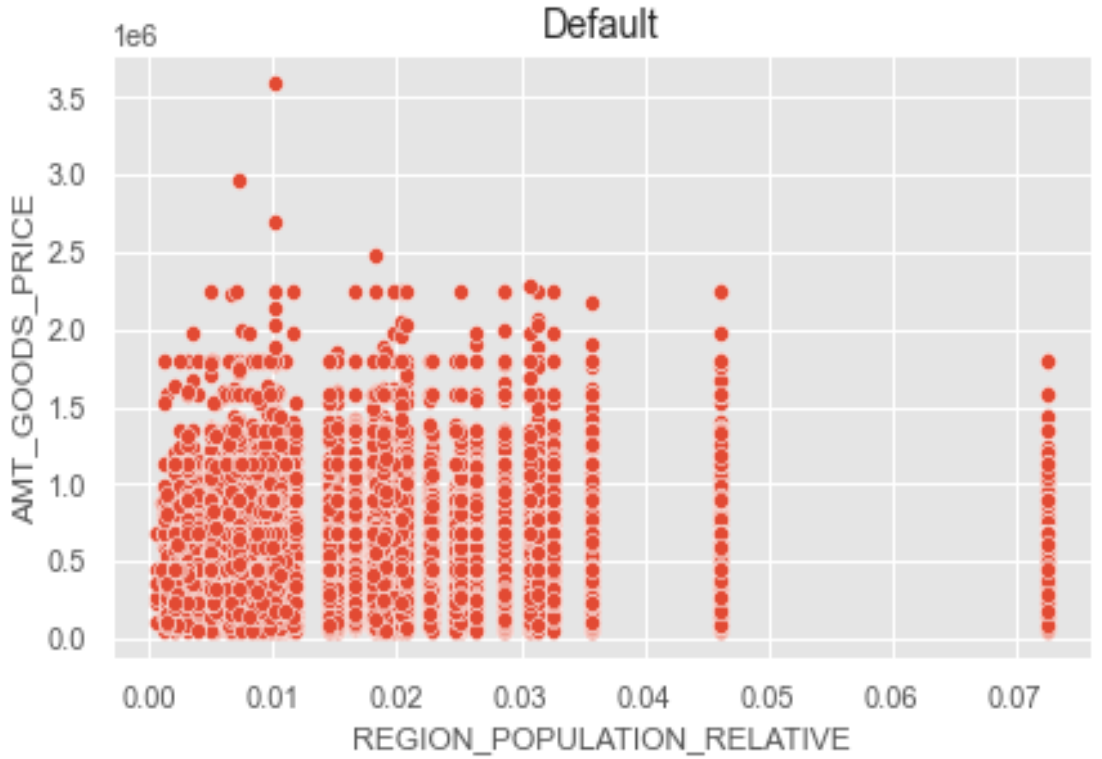




Goods price vs Region population relative

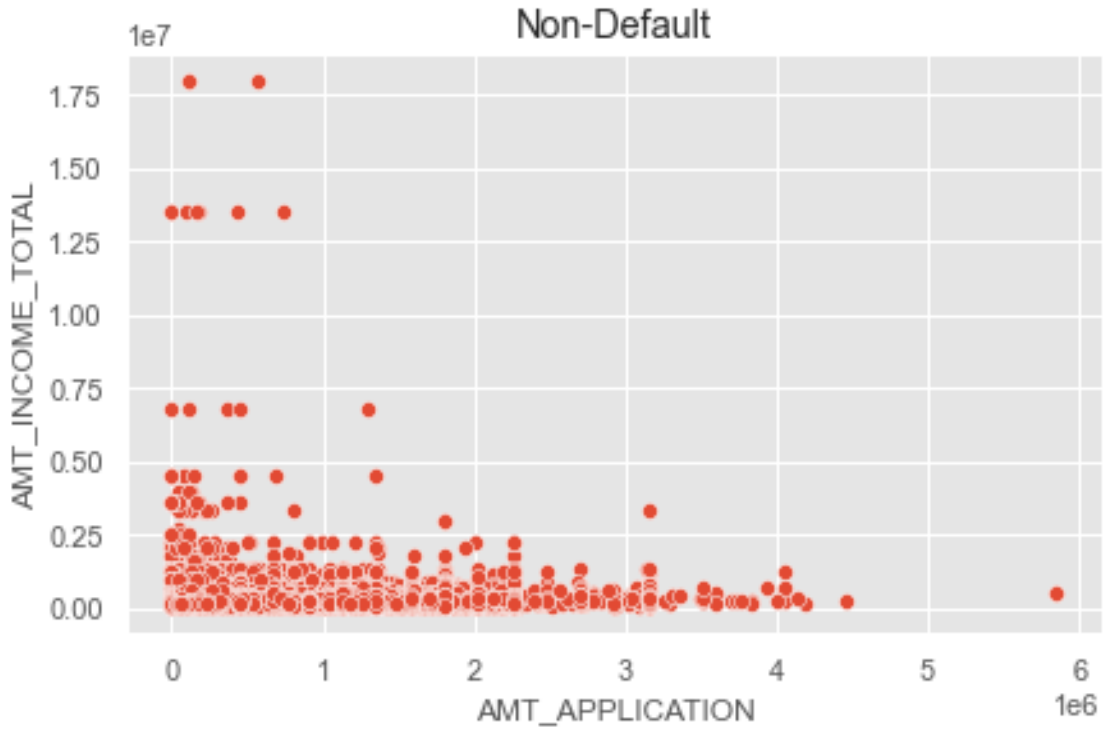
- Defaulters live in less populated region & applied for Lesser Goods Amount Credit when compared to Non-Defaulters.

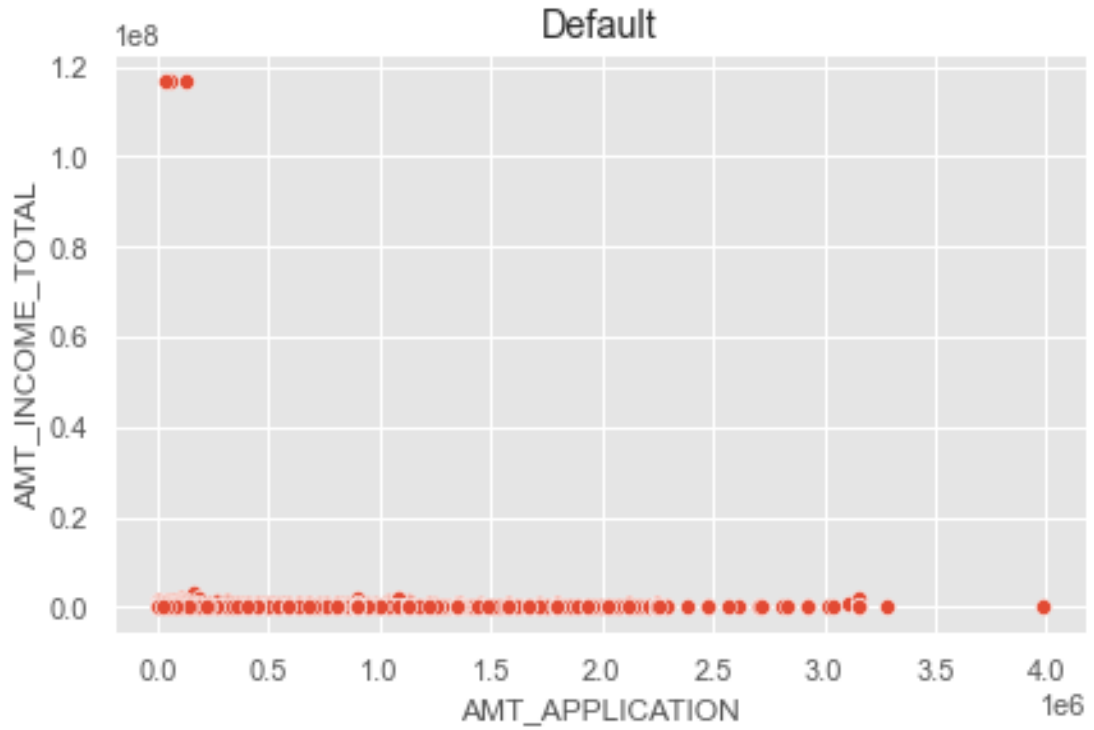




Amount Application vs Amount Income total

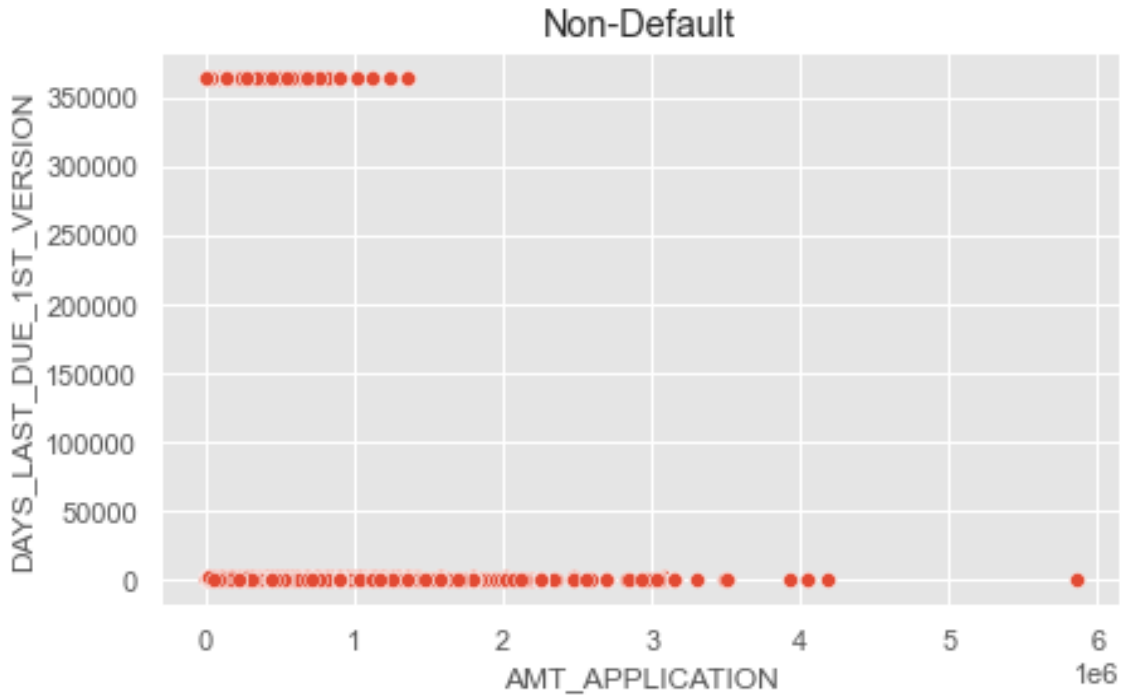
- In Non Defaulters, the data points are clustered at the low level and scattered on increasing points and observed exponential decrease between these variables.
- Defaulters having low income have applied for lower credit amounts compared to the Non-Defaulters.

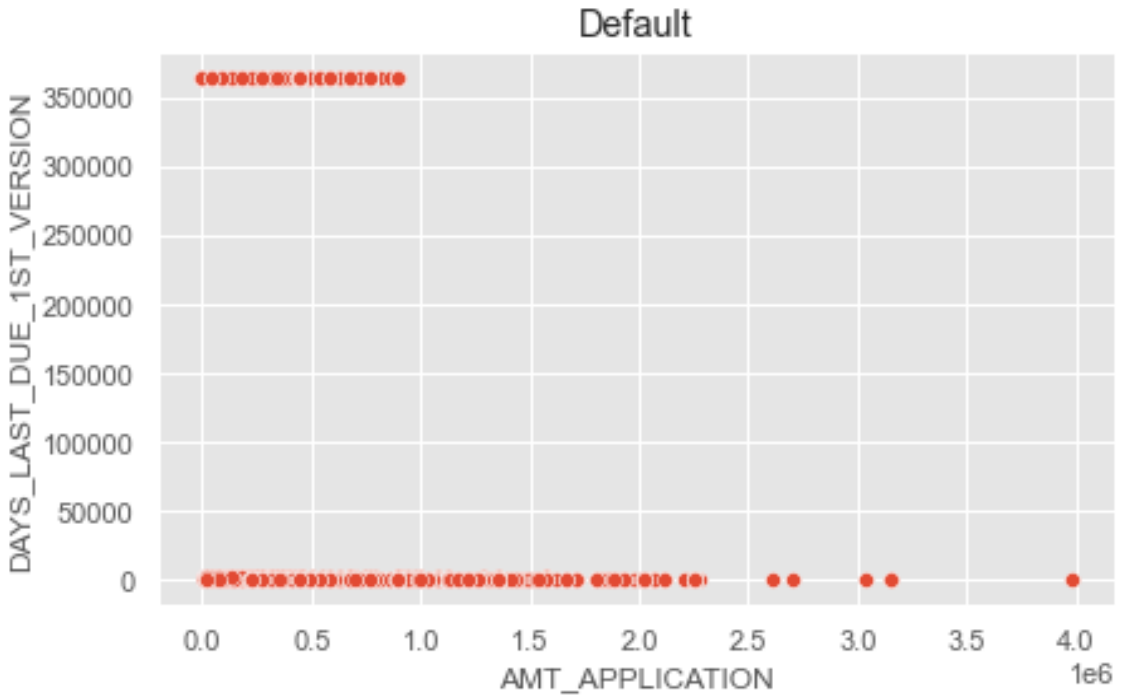




AMT-APPLICATION Vs DAYS-LAST-DUE-1ST-VERSION

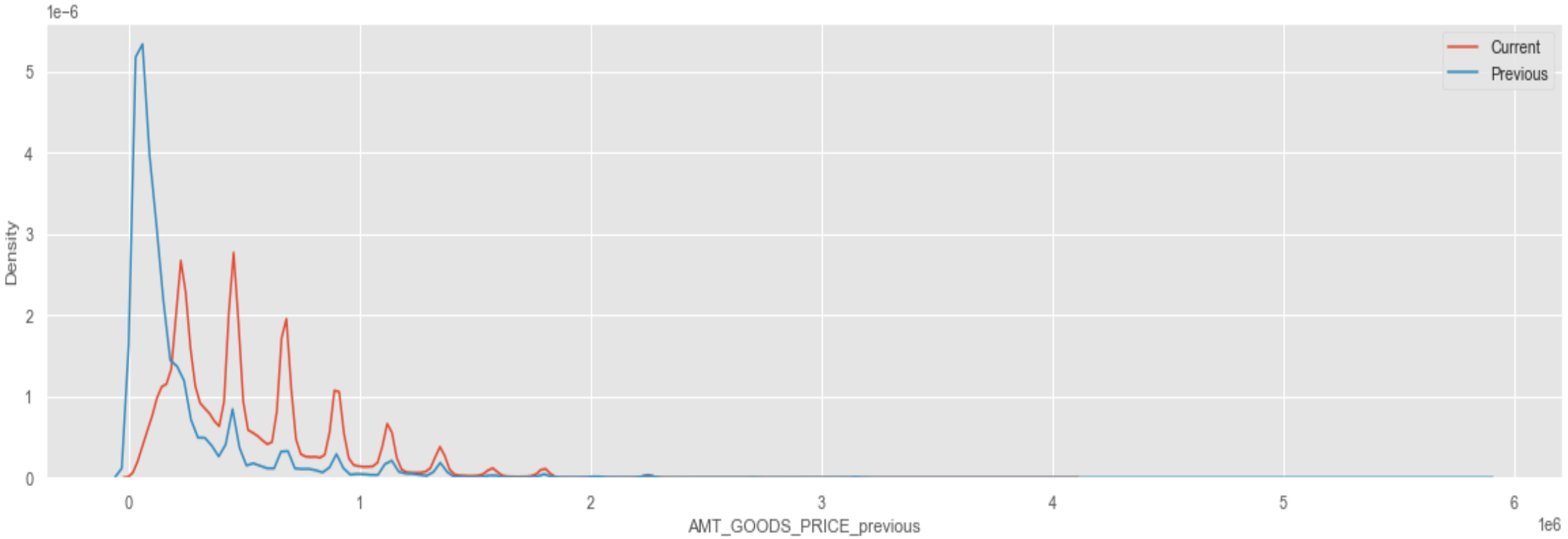
- Defaulters Application amount was less compared to Non-Defaulters application amount.





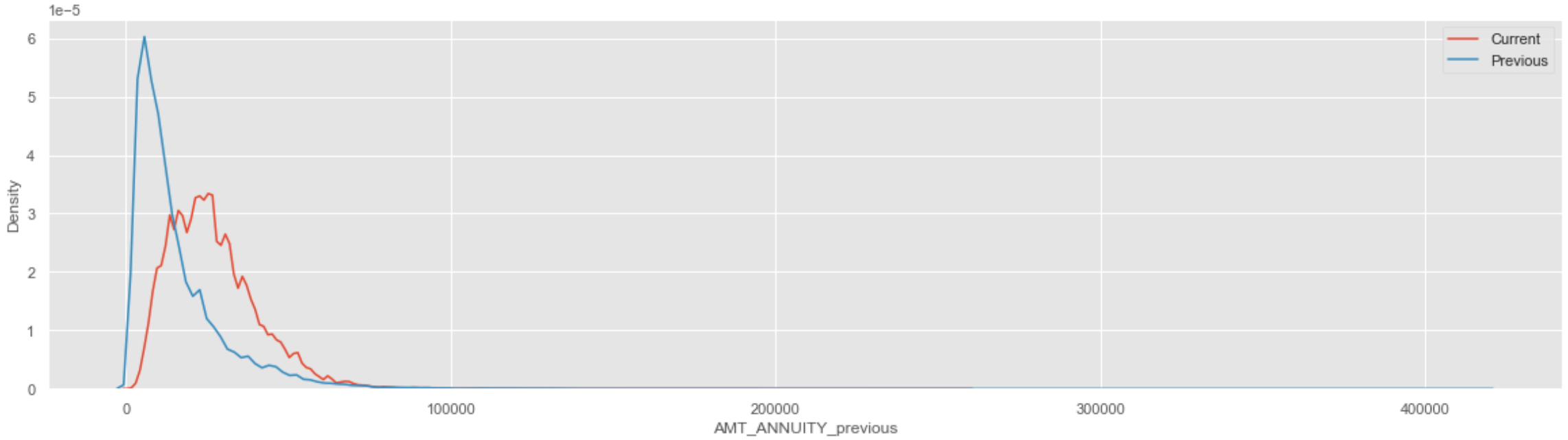
Amount Goods Price Current vs Amount Goods Price Previous

- Total Amt_Goods_Previous seems to be higher than Amt_Goods_current.



Amount Annuity current vs Amount Annuity Previous

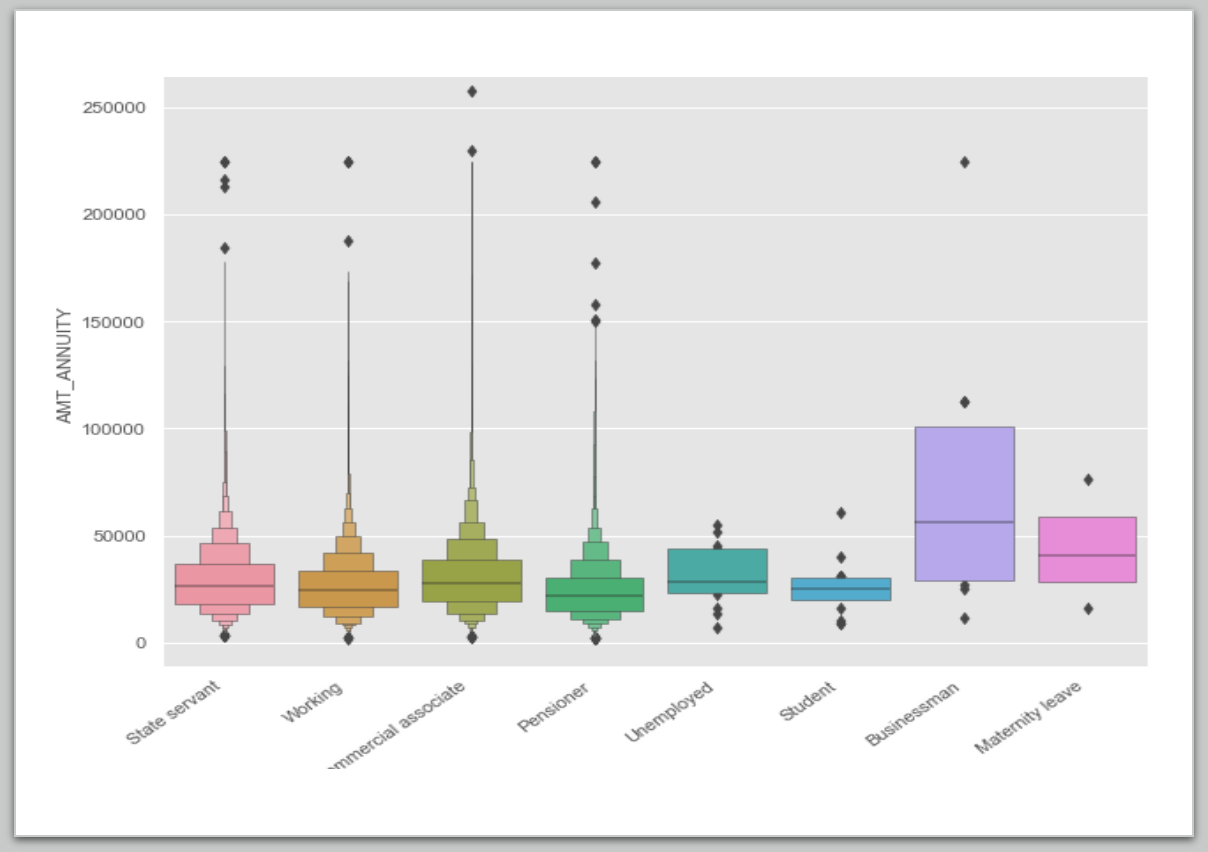
- Total Amt_Annuity_previous is higher than the Amt_Annuity_current.

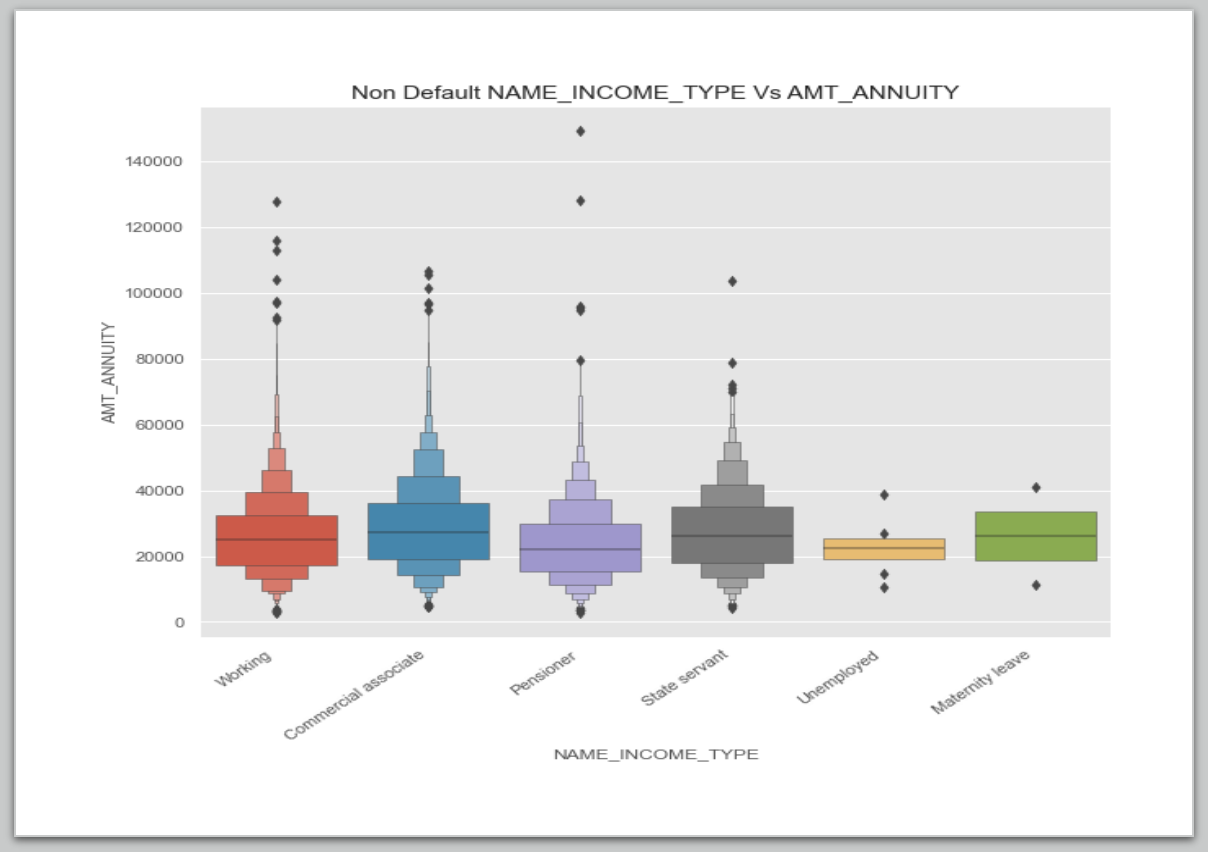


Bivariate Analysis Categorical - Numerical Application Dataset.

Amount Annuity vs Name Income Type

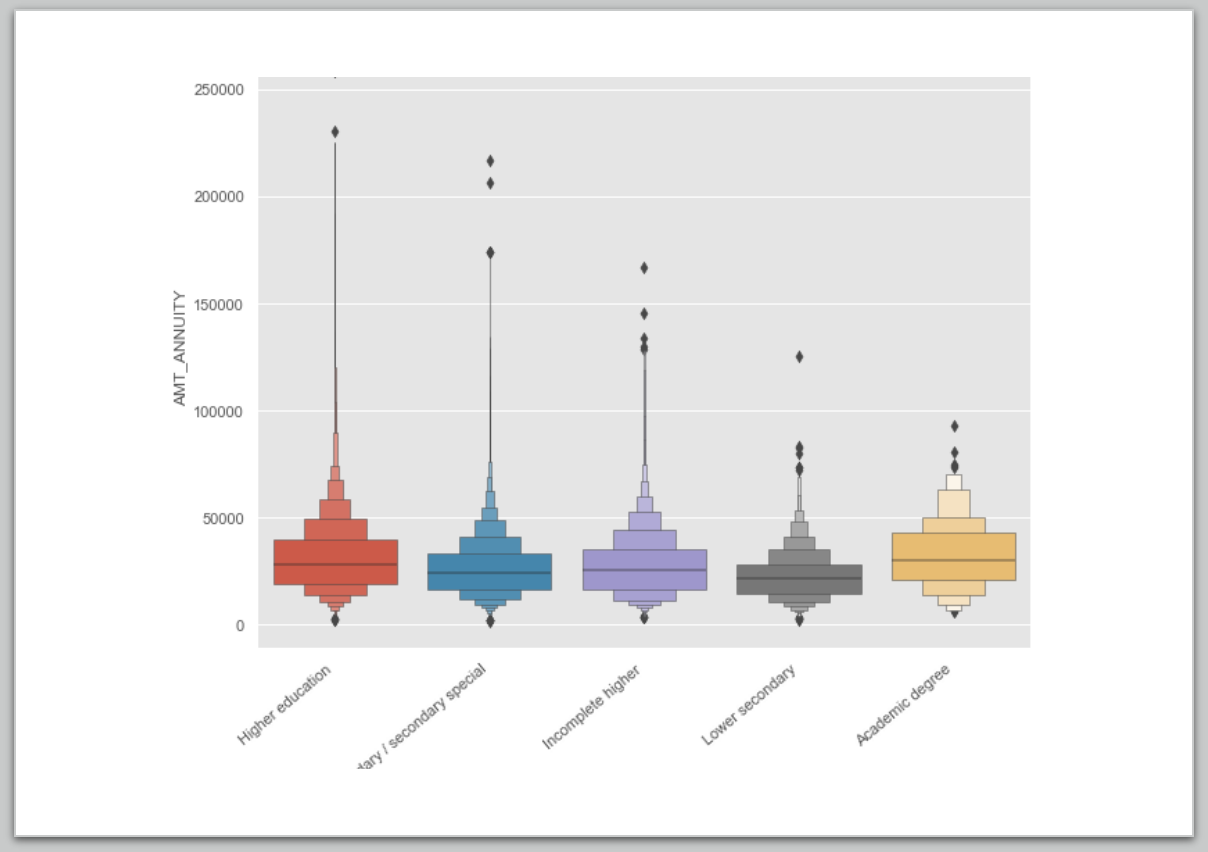
- Among Defaulters, Businessman pay the highest Annuity & Students pay the least Annuity.
- Among Non- Defaulters, Commercial Associates pay the highest Annuity & Unemployed pay the least Annuity.
- There appear many data points above the Upper Whisker & Lower Whisker, but cannot be considered as Outliers unless we look into the Credit amounts.

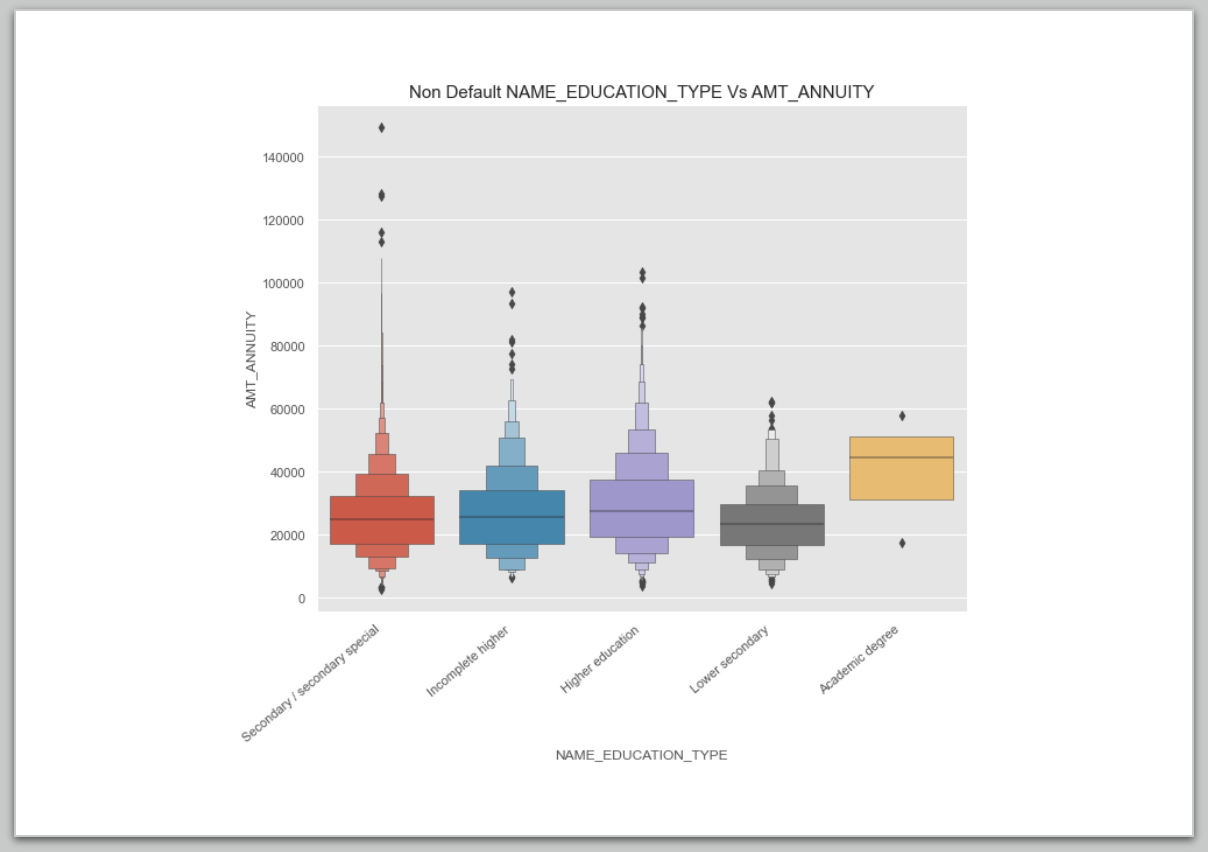




NAME_EDUCATION_TYPE vs AMT_ANNUITY

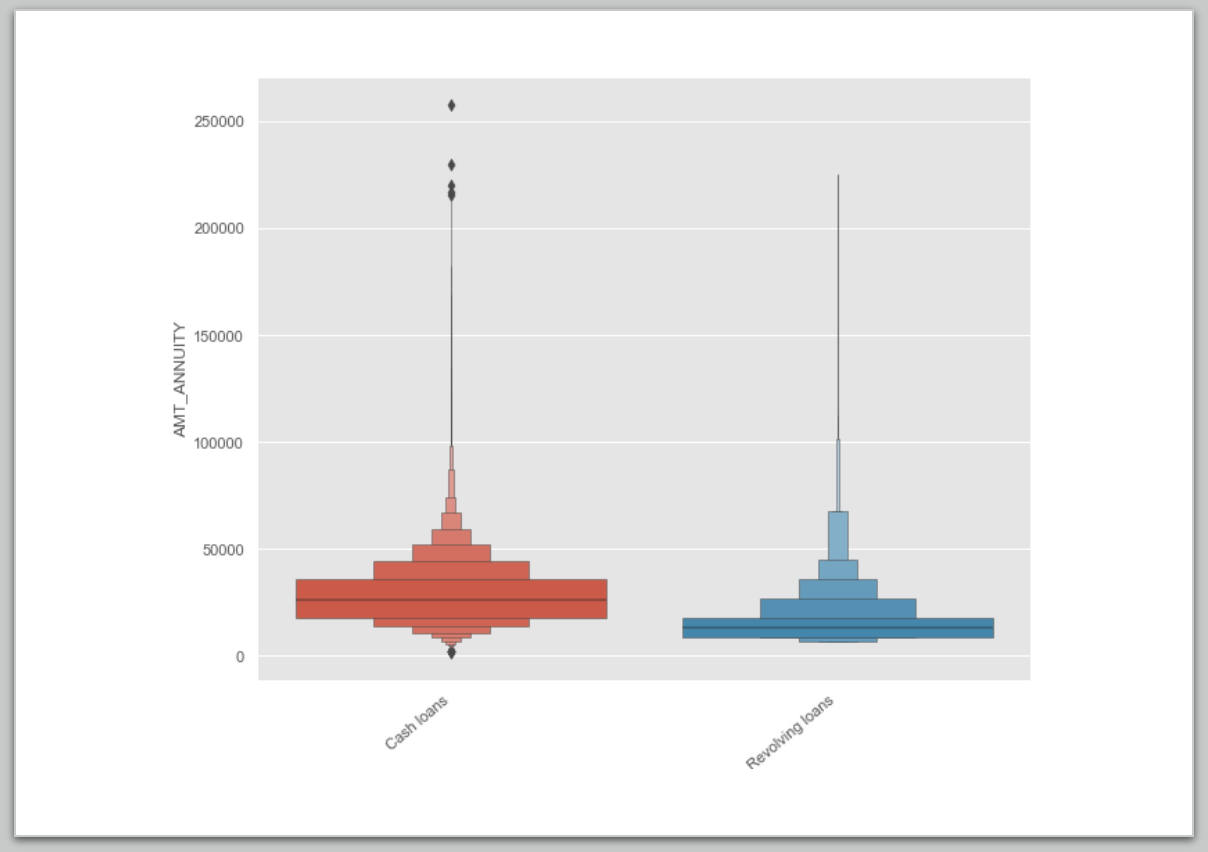
- Highest no of Defaulters have Higher Education & defaulted Annuity between 2- 2.5 Lakhs
- Highest no of Non-Defaulters have Higher Education & pay High Annuity between 1-1.2 Lakhs

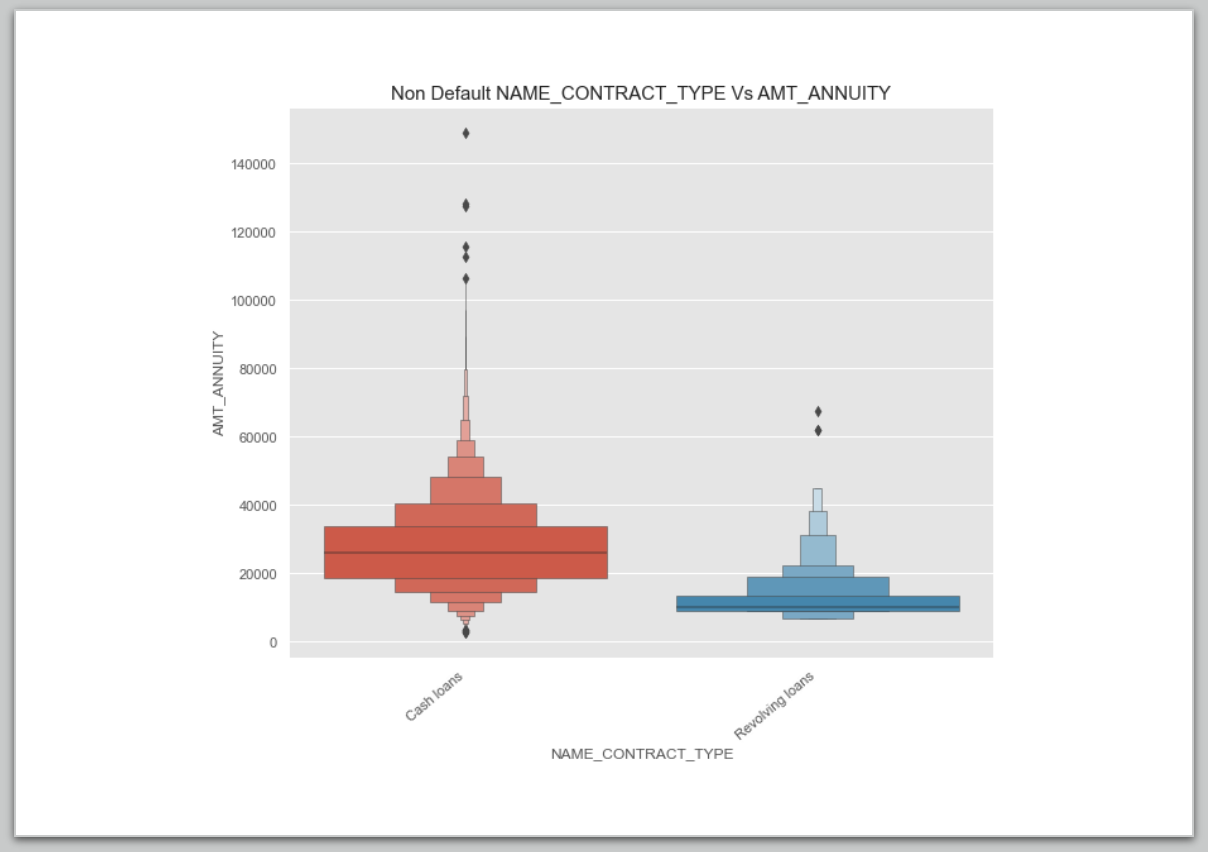




NAME_CONTRACT_TYPE vs AMT_ANNUITY

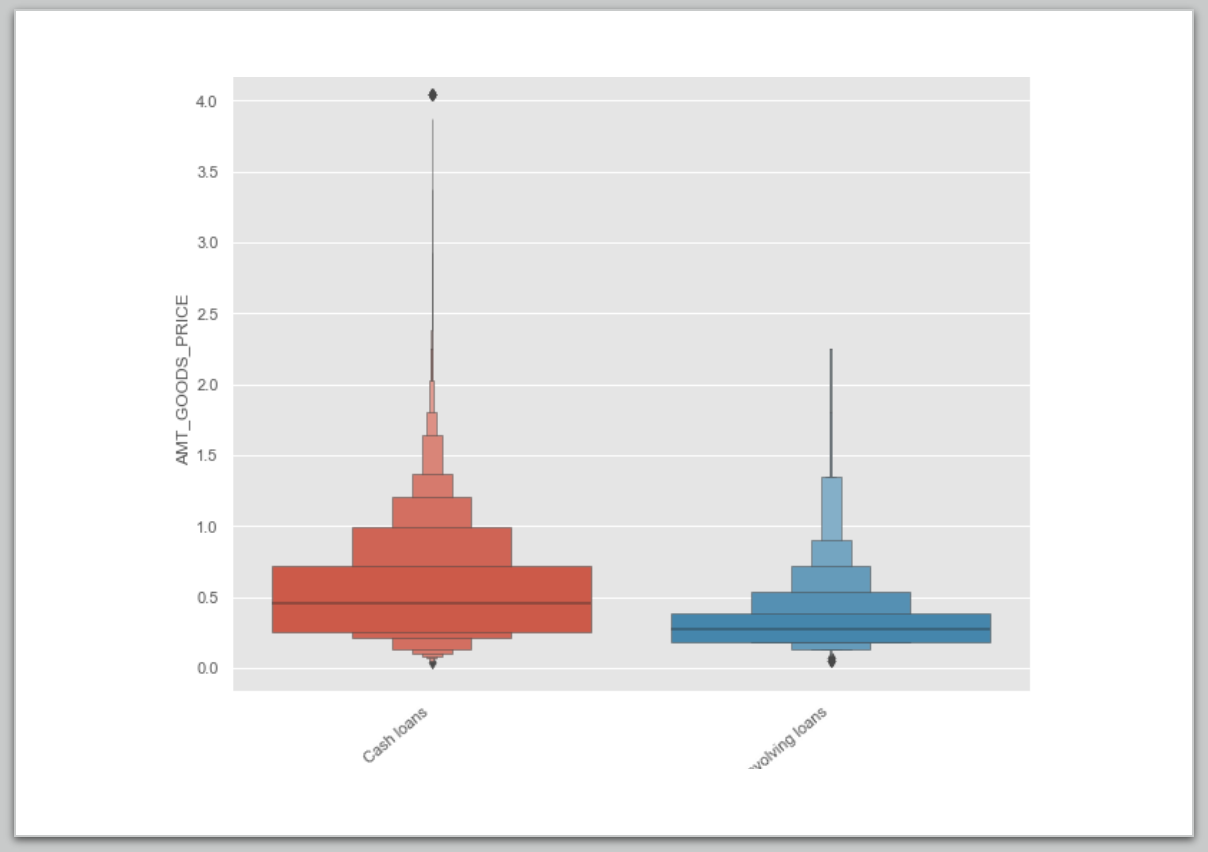
- Most of the Defaulters have opted for Cash Loans with Annuity between 2-2.5 Lakhs.
- Non-Defaulters have opted for Cash Loans with Annuity of 1-1.2 Lakhs.

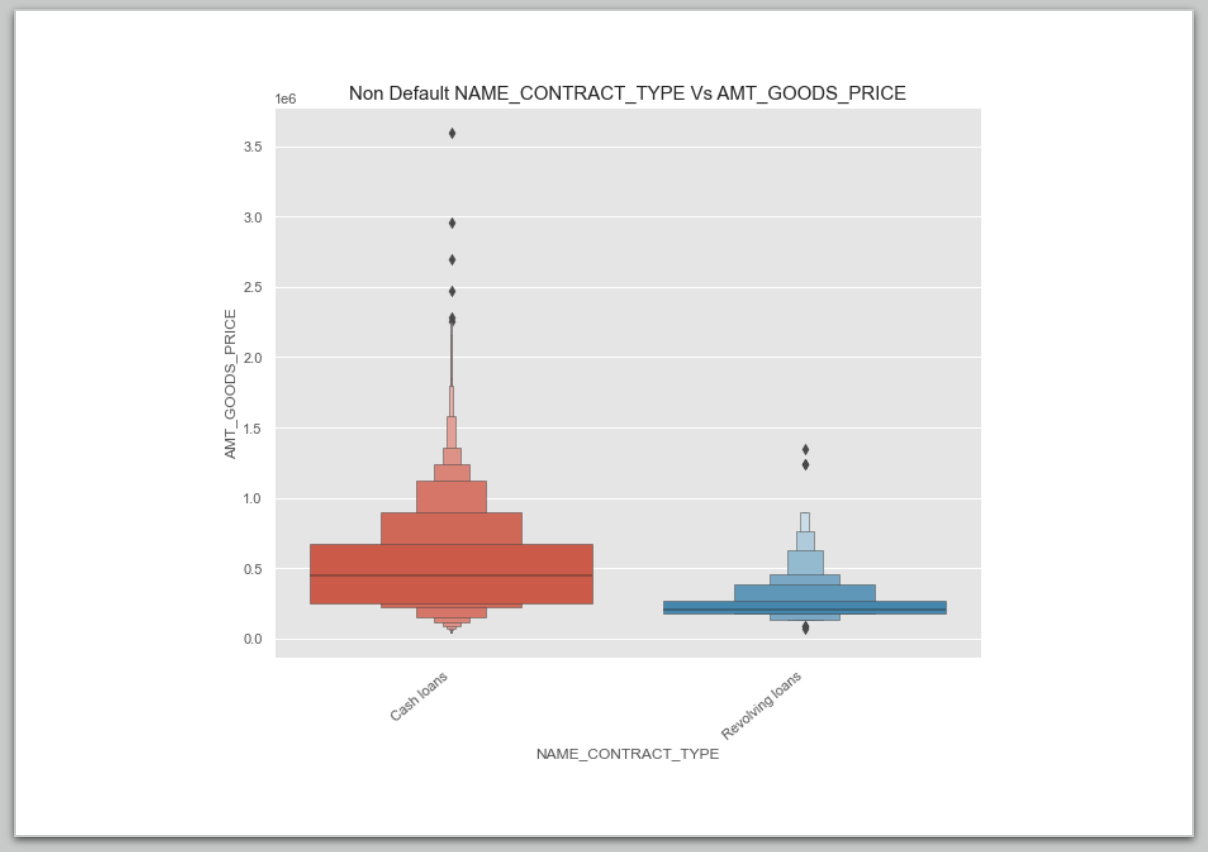




NAME_CONTRACT_TYPE vs AMT_GOODS_PRICE

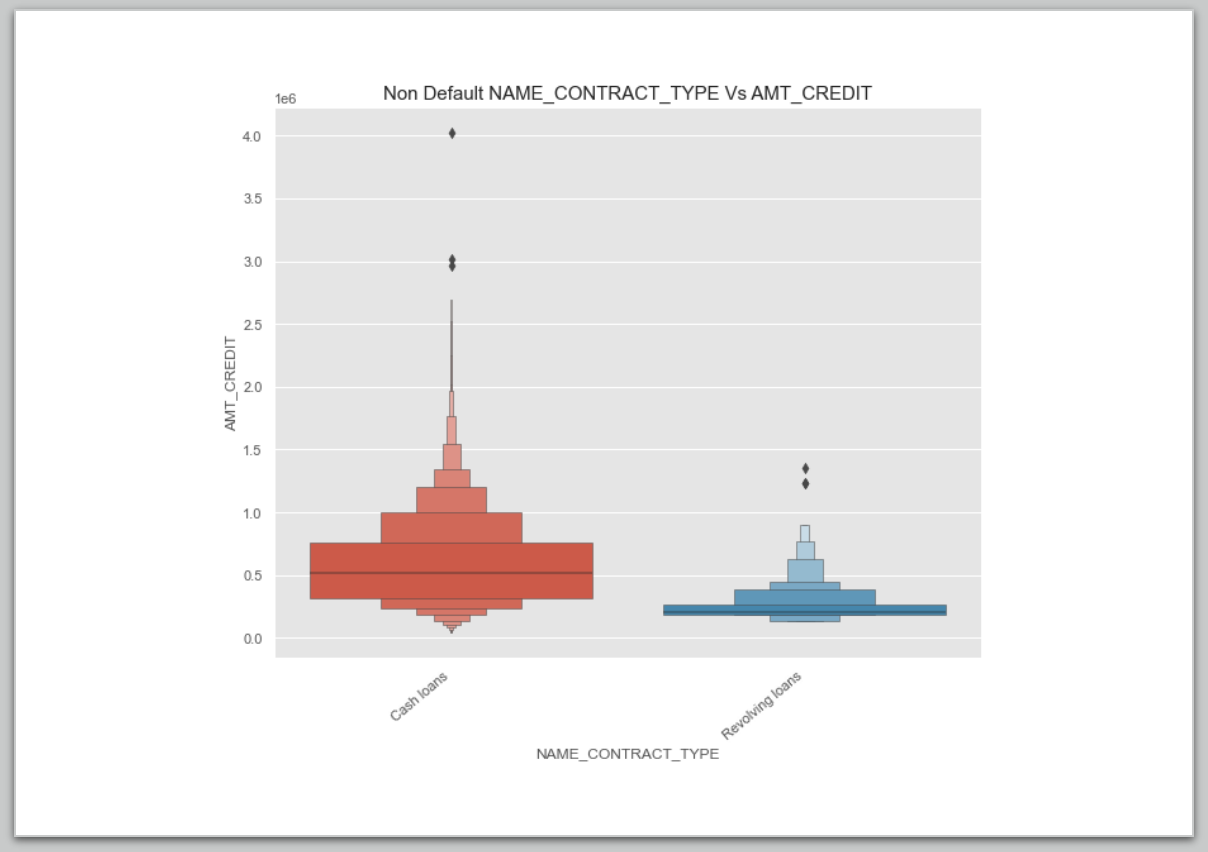
- Most of the Defaulters Opted Cash Loans for Goods Price.
- Most of the Non-Defaulters opted for Cash loans for Goods Price

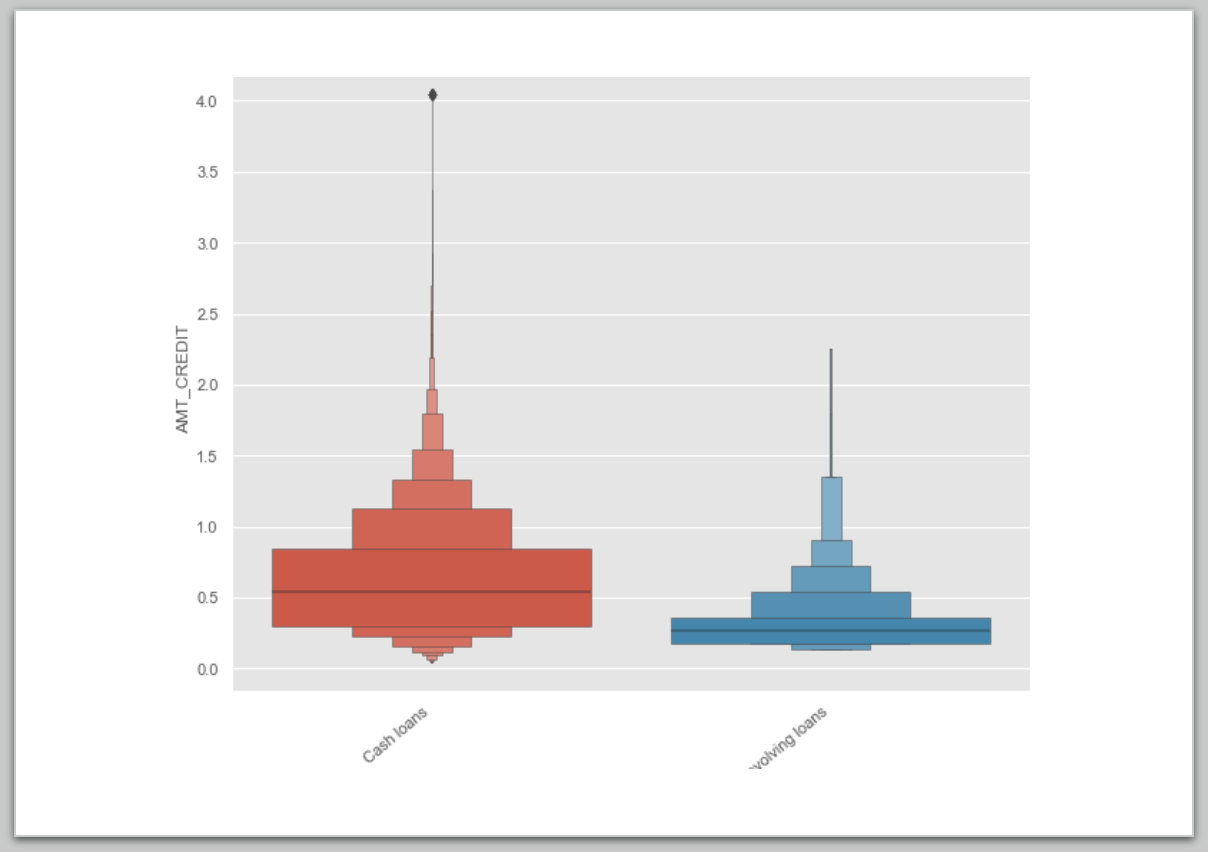




NAME_CONTRACT_TYPE vs AMT_CREDIT

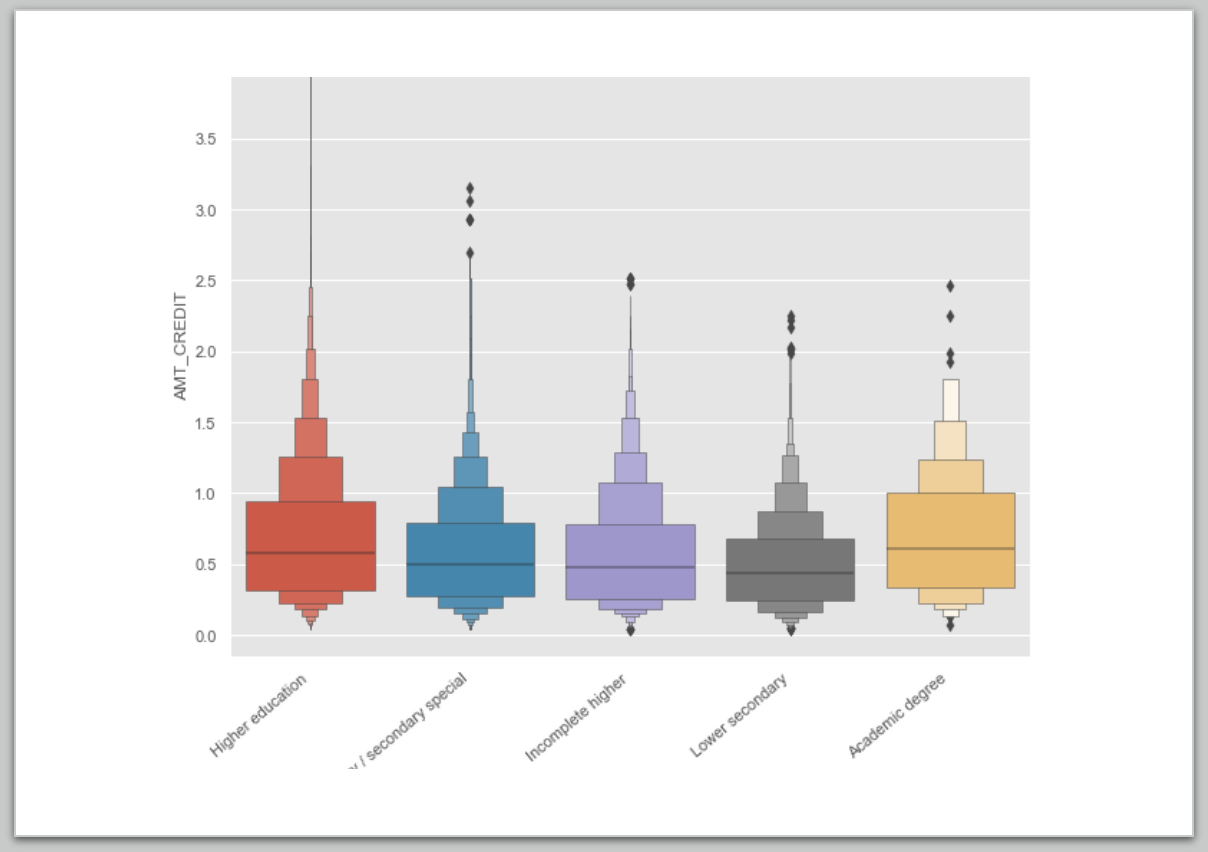
- Most of the Defaulters & Non-Defaulters opted for Cash loans over Revolving Loans.

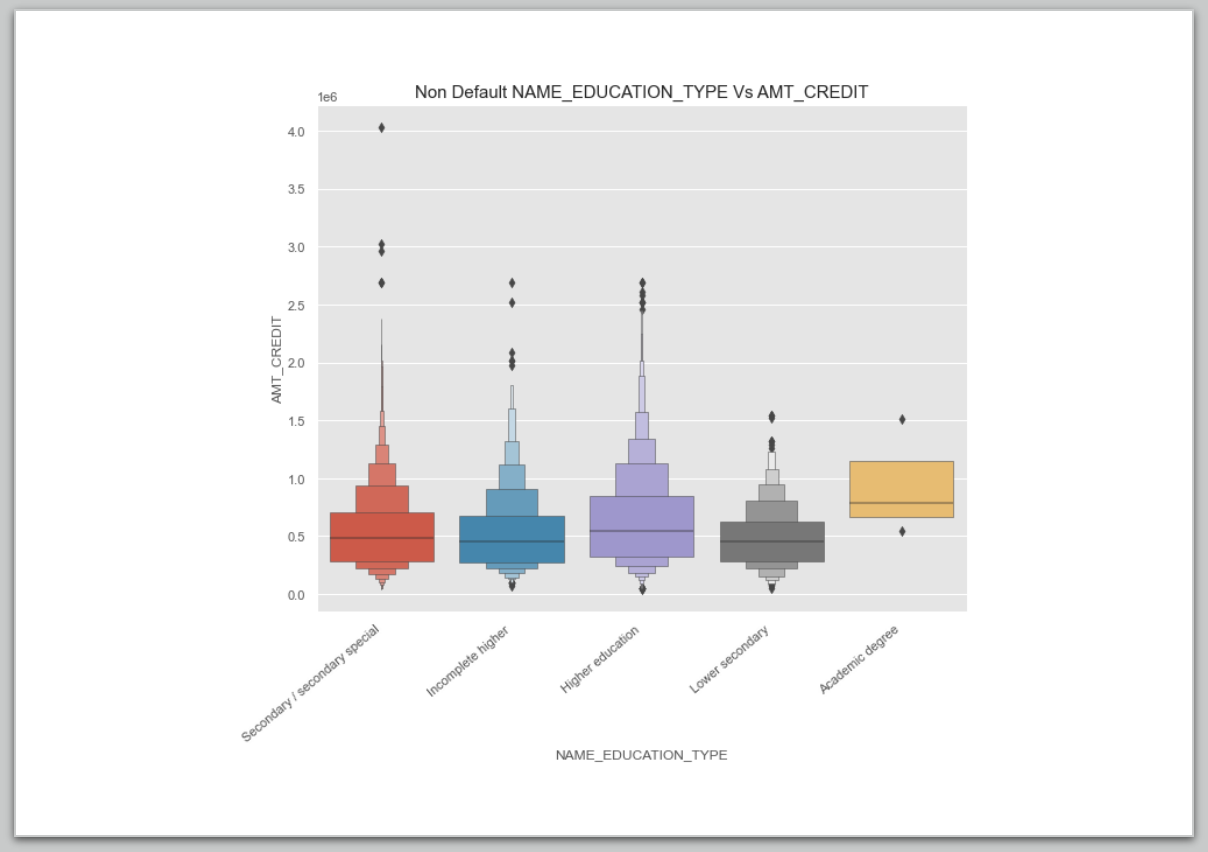




NAME_EDUCATION_TYPE vs AMT_CREDIT

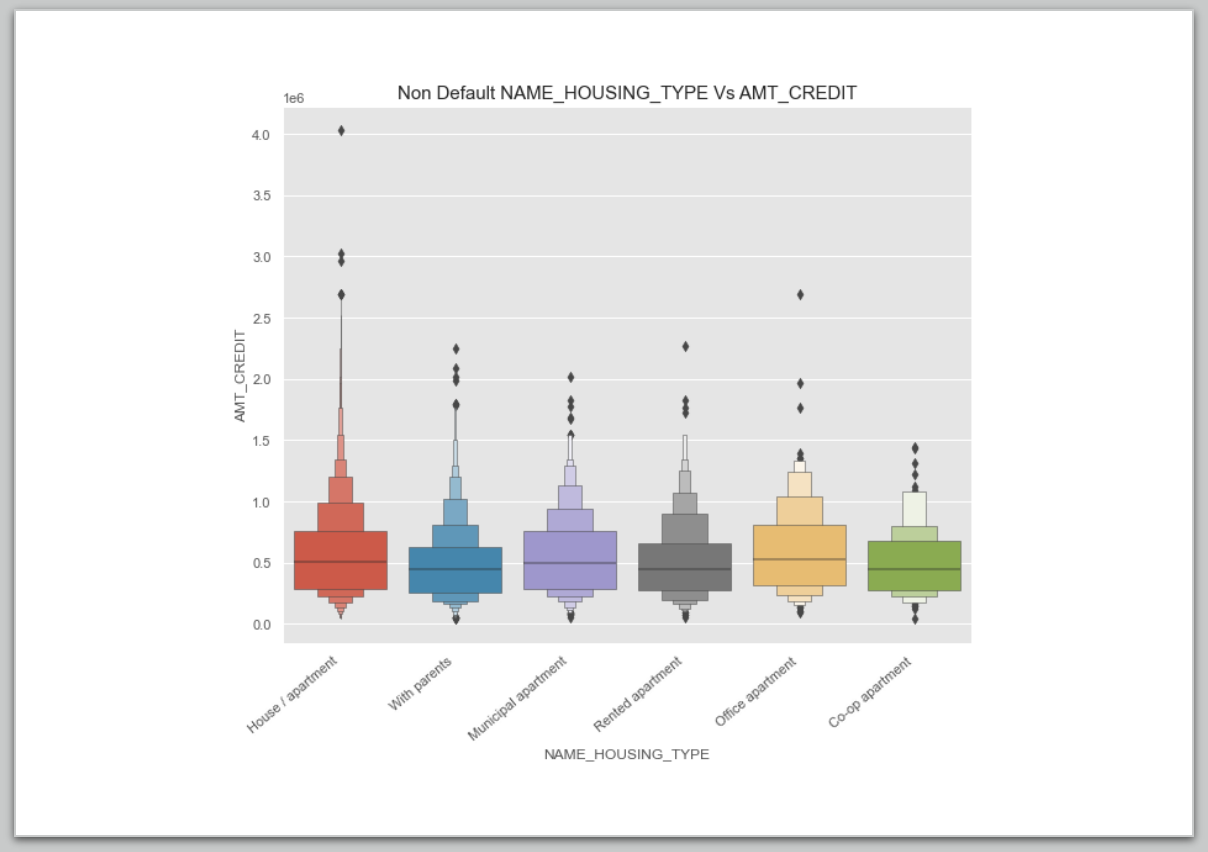
- Most of the Defaulters with Higher Education got the Credit approved.
- Most of the Non-Defaulters with Higher Education got Credit approved.

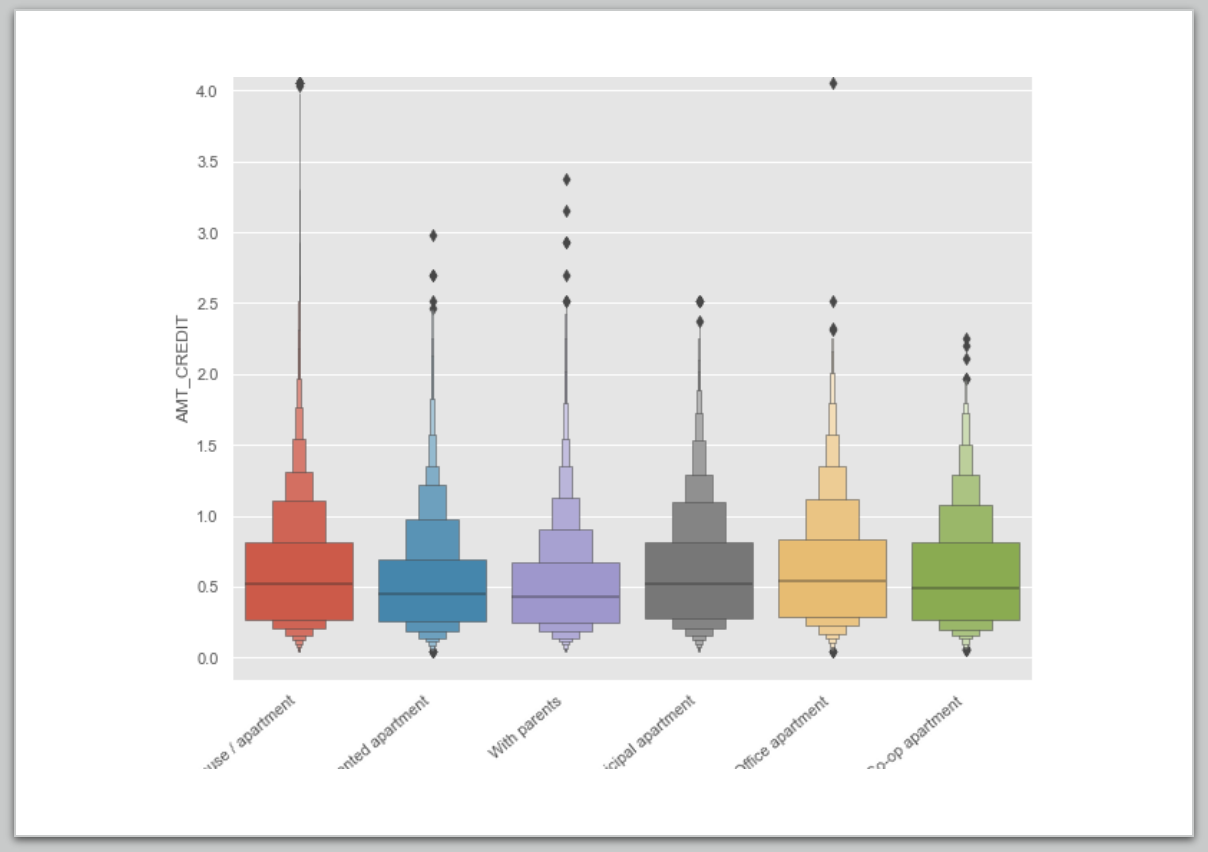




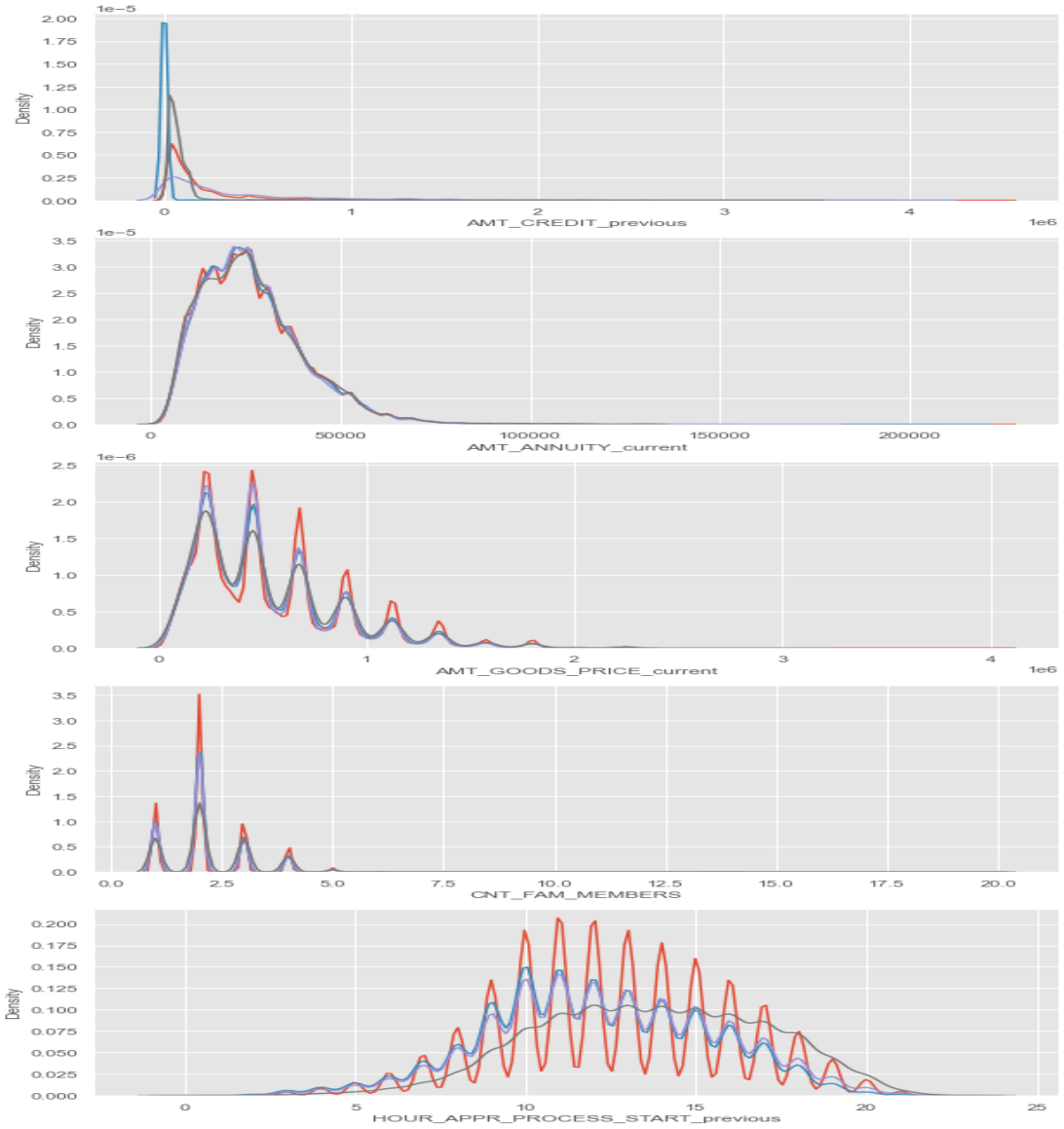
NAME_HOUSING_TYPE vs AMT_CREDIT

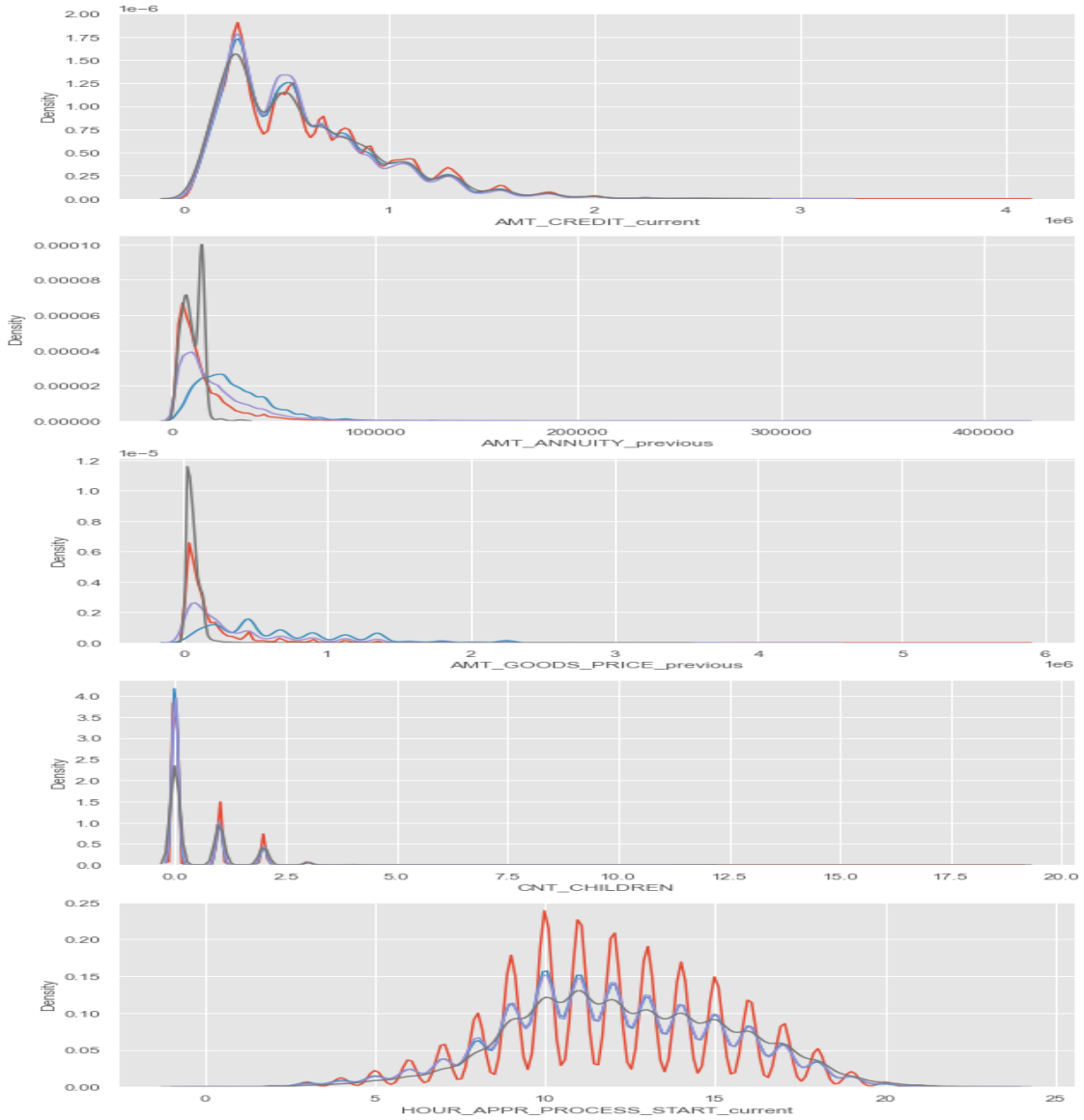
- Most of the Defaulters who own House/Apartment got the Credit approved.
- Most of the Non-Defaulters who own House/Apartment have for Credit approved.





Categorical to Numerical Analysis





Categorical to Numerical Analysis(contd..)

Observations from the previous plots:

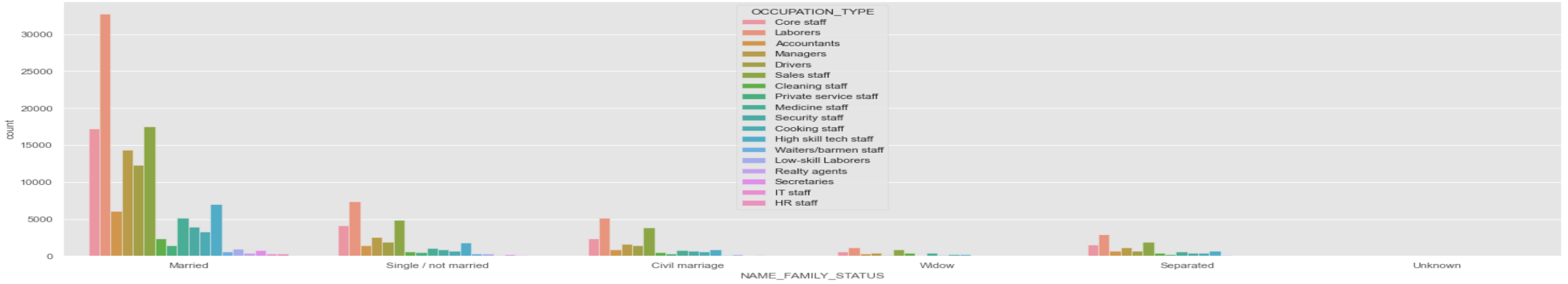
- High number of applications are filed in 9 AM to 2 PM for both Current and Previous data. So busiest hours for bank are from 9 AM to 2 PM.
- Nuclear family tends to take more loans.
- Previously bank had high unused offers but currently refused is high incase of AMT_GOODS_PRICE.
- Previously bank had high unused offers and currently cancelled/refused offers are similar for AMT_ANNUITY.
- Previously bank had high unused offers and currently high number of refused offers for AMT_CREDIT.

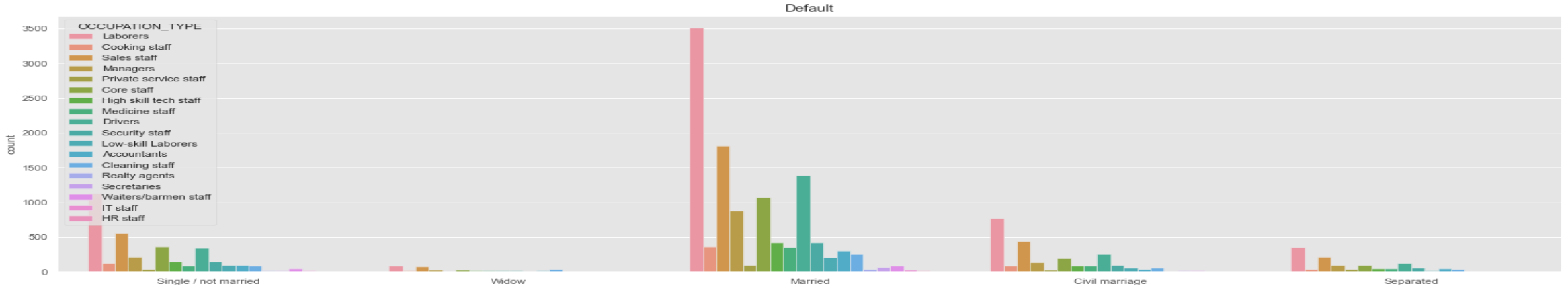
Bivariate Analysis

Categorical to Categorical Analysis

Family Status vs Occupation type

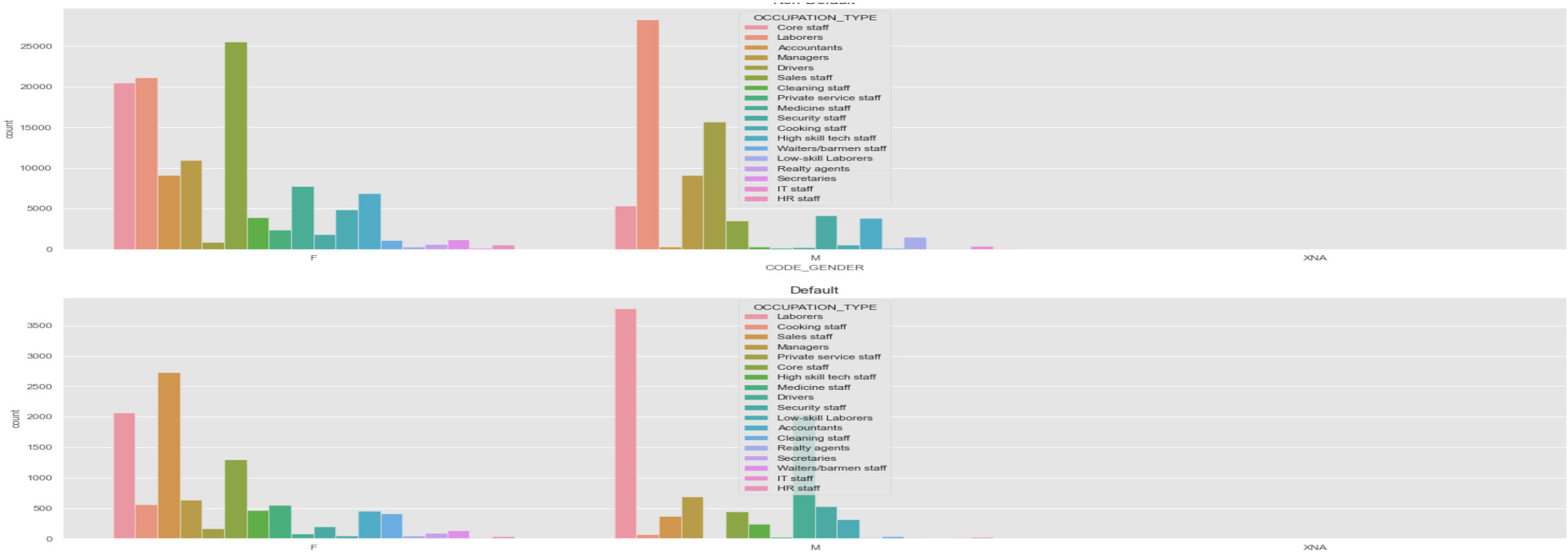
- Most of the Non-Defaulters are Married & Core Staff by occupation. Least are Widow with Laborer occupation.
- Most of the Defaulters are Married & Laborers by Occupation and least are Widow & Laborers.





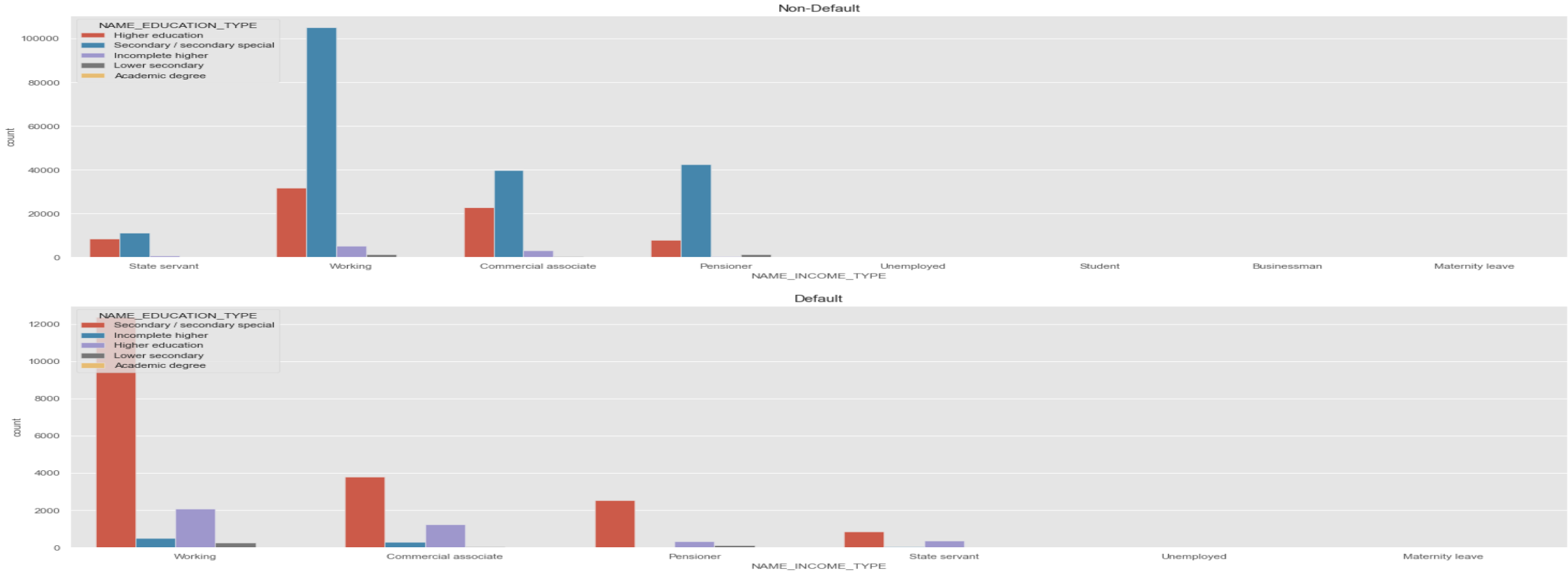
Code Gender vs Occupation Type

- Most of the Non-Defaulters are Male & Laborers and least is Male in IT Staff
- Most of the Defaulters are Male & Laborers and least is Male in IT Staff



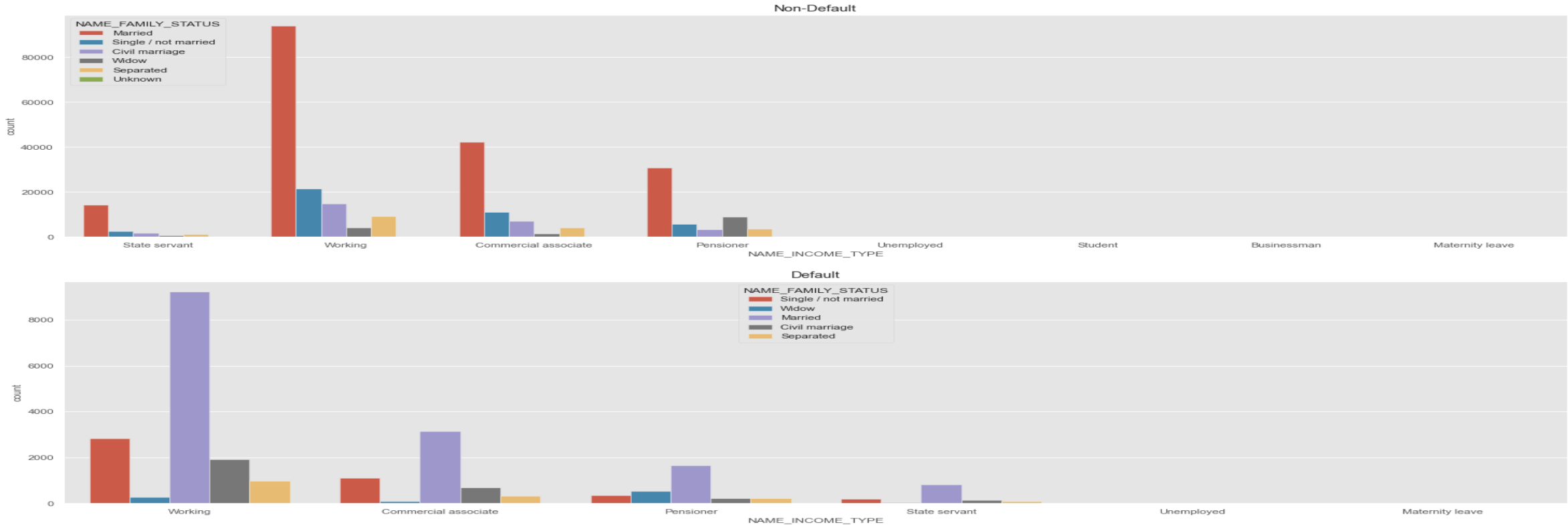
Income Type Vs Education Type

- Most of the Non-Defaulters are Working with Secondary/ Secondary Special education.
- Most of the Defaulters are Working with Secondary/Secondary Special education.



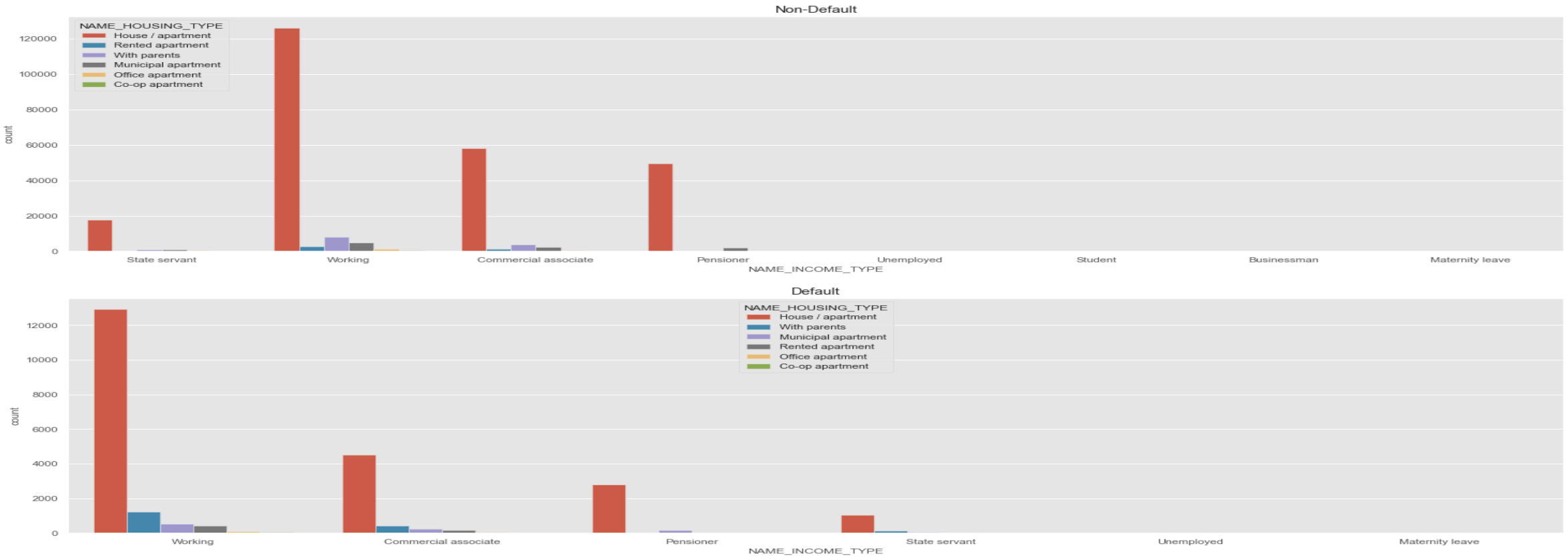
Income Type vs Family Status

- Most of the Non- Defaulters are Working & Married.
- Most of the Defaulters are Working & Married.



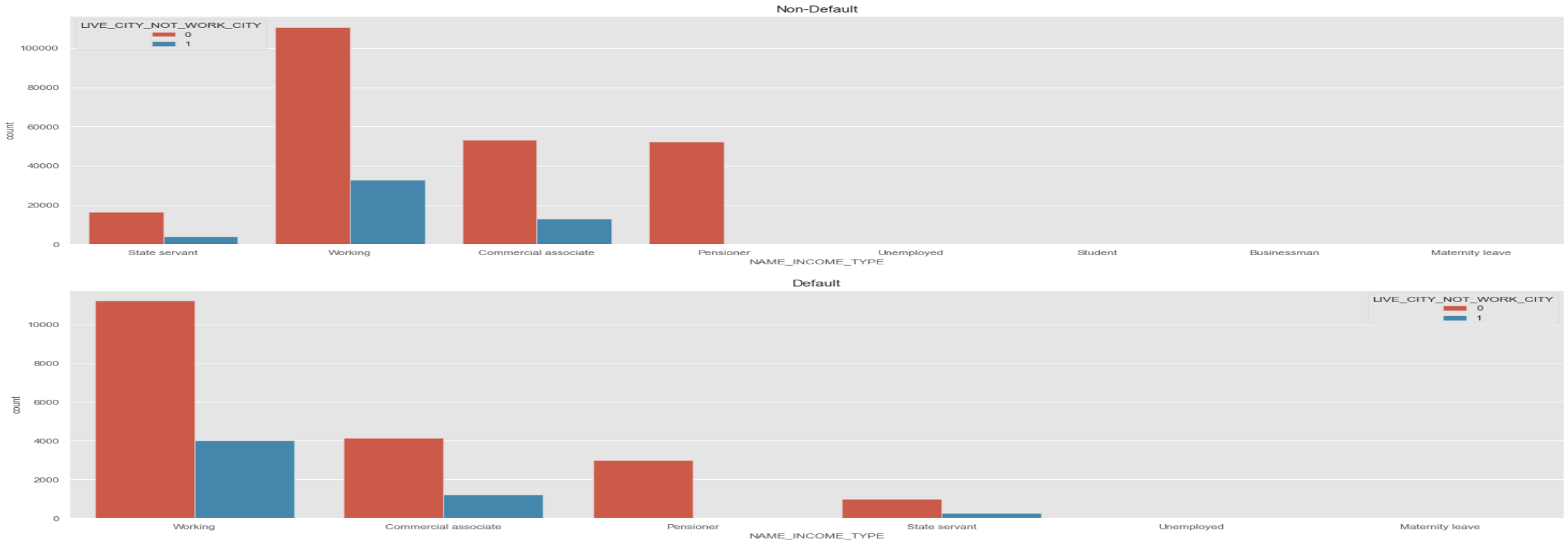
Income Type vs Name housing Type

- Most of the Non- Defaulters are Working & Own House/Apartment.
- Most of the Defaulters are Working & own a House/Apartment, least are State Servant & stay with Parents.



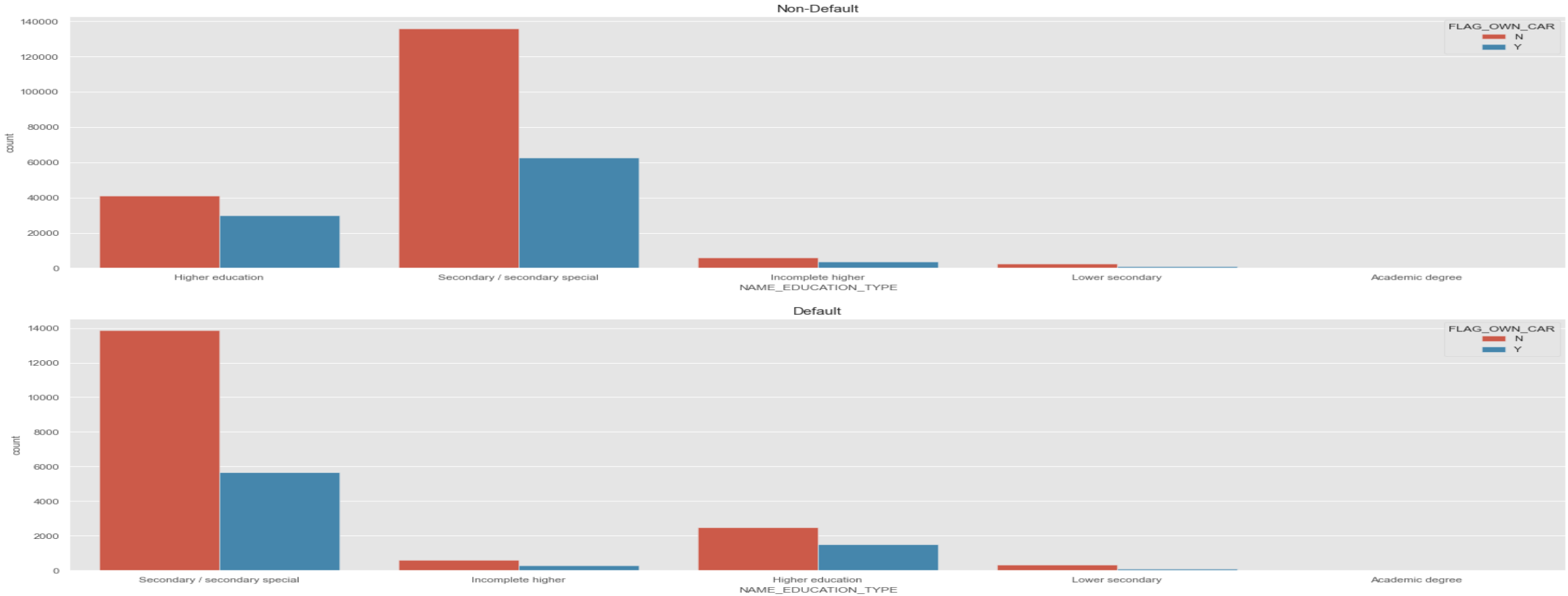
Income Type vs LIVE_CITY_NOT_WORK_CITY

- Most of the Non-Defaulters Contact address match with the work City.
- Most of the Defaulters Contact address match with the work City.



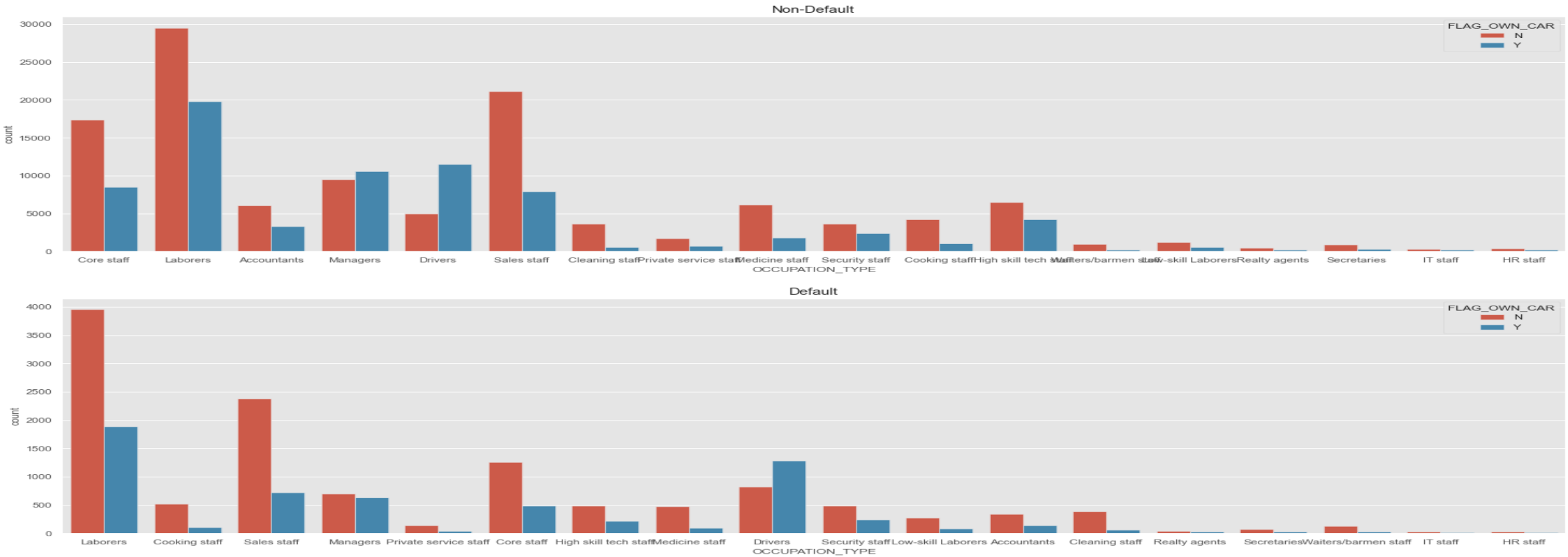
Education Type vs OWN CAR

- Most of the Non-Defaulters & Defaulters with Secondary/Secondary Special & do not own a car.



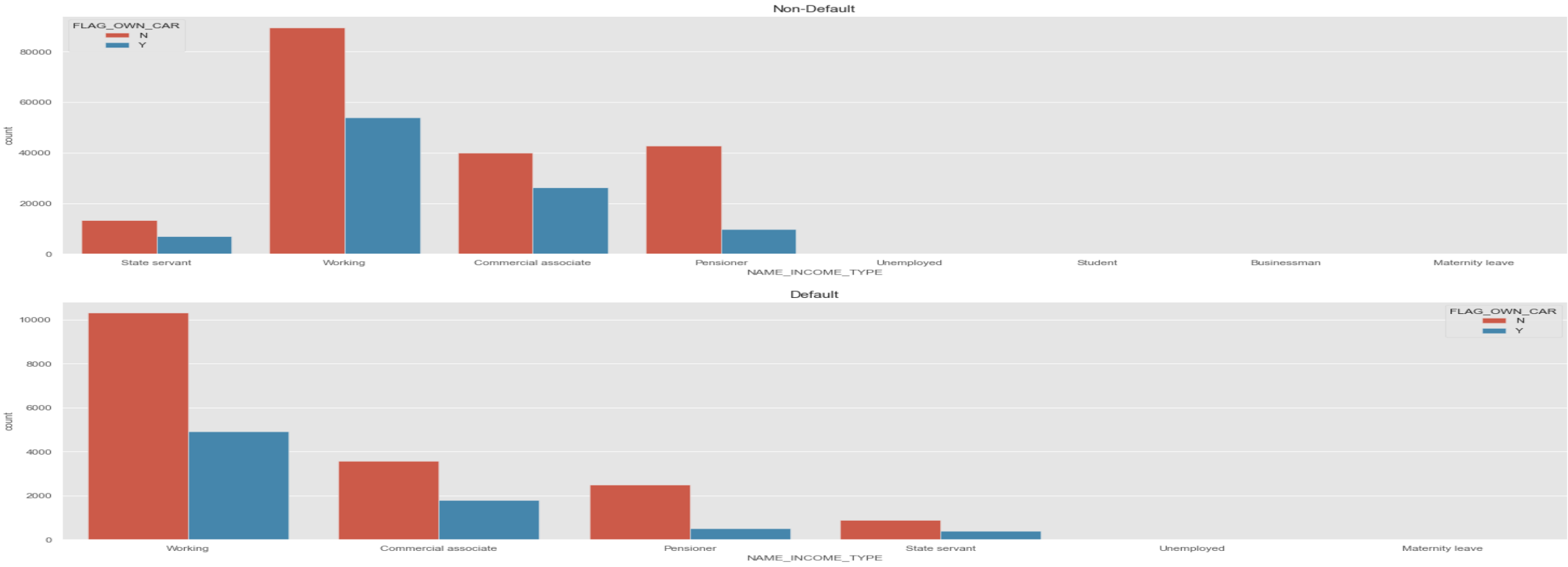
Occupation Type vs Own Car

- Most of the Non-Defaulters are Laborers & don't own a Car.
- Most of the Defaulters are Laborers & don't own a Car



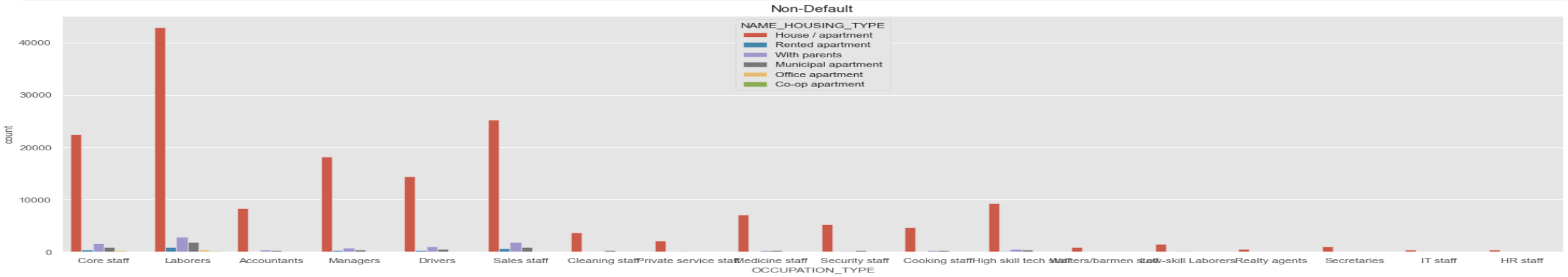
Name Income Type vs Flag Own Car

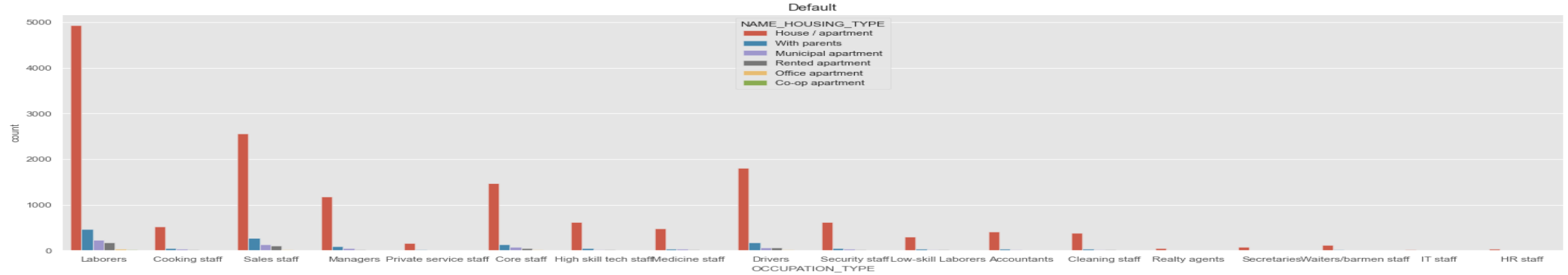
- Most of the Non-Defaulters are Laborers & don't own a Car.
- Most of the Defaulters are Laborers & don't own a Car.



Occupation Type vs NAME HOUSING TYPE

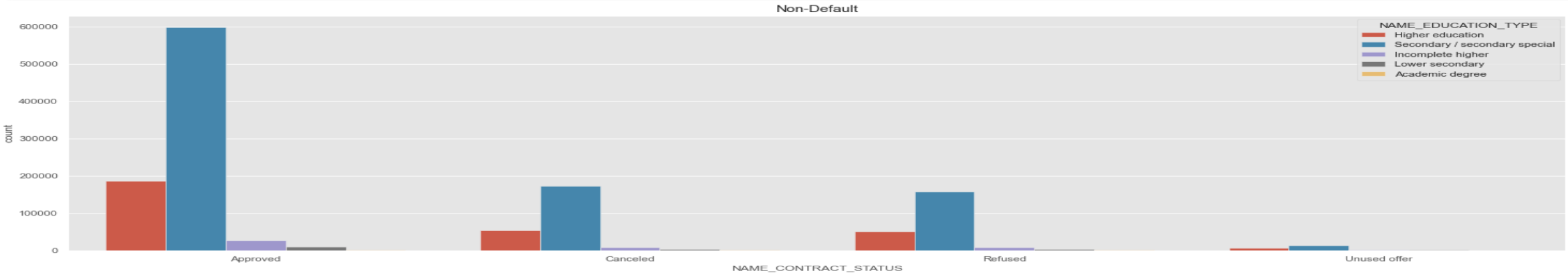
- Most of the Non-Defaulters are Laborers with own House/Apartment & least are the HR Staff with Own House/Apartment.
- Most of the Defaulters are Laborers with own House/Apartment & least are the IT Staff with own House/Apartment

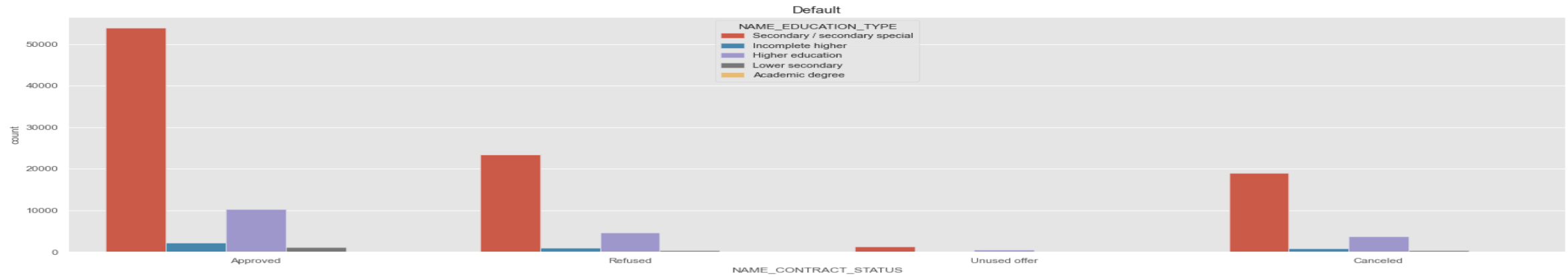




Name Contract Status vs Name Education Type

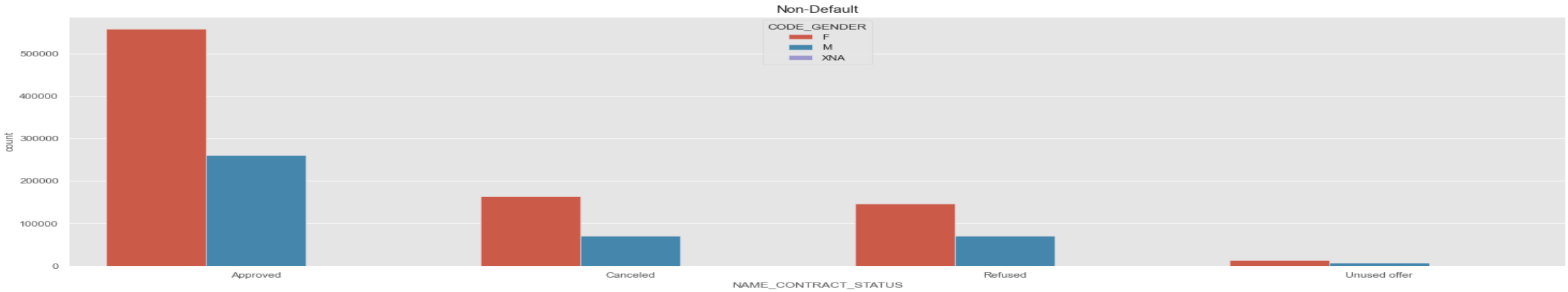
- Most of the Approved Credit & Non-Defaulters Education is Secondary/Secondary Special.
- Most of the Approved Credit & Defaulters Education is also Secondary/Secondary Special.

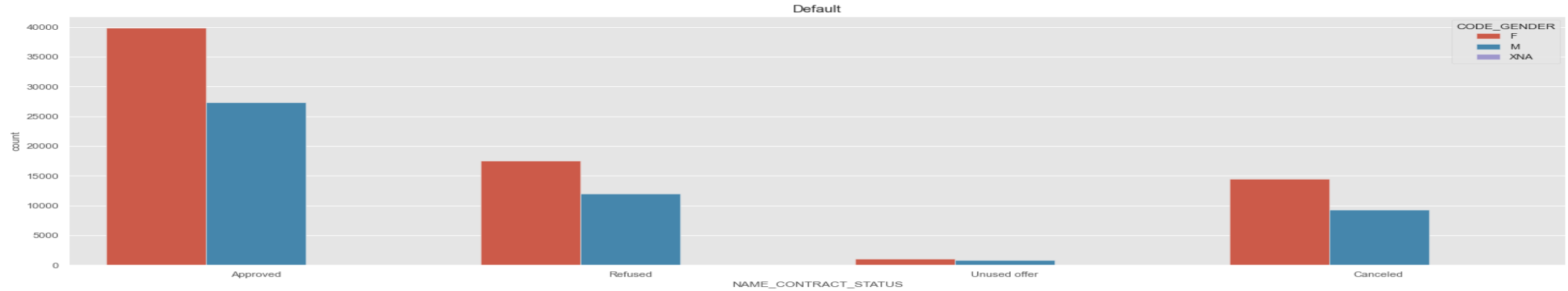




Name Contract status vs Code Gender

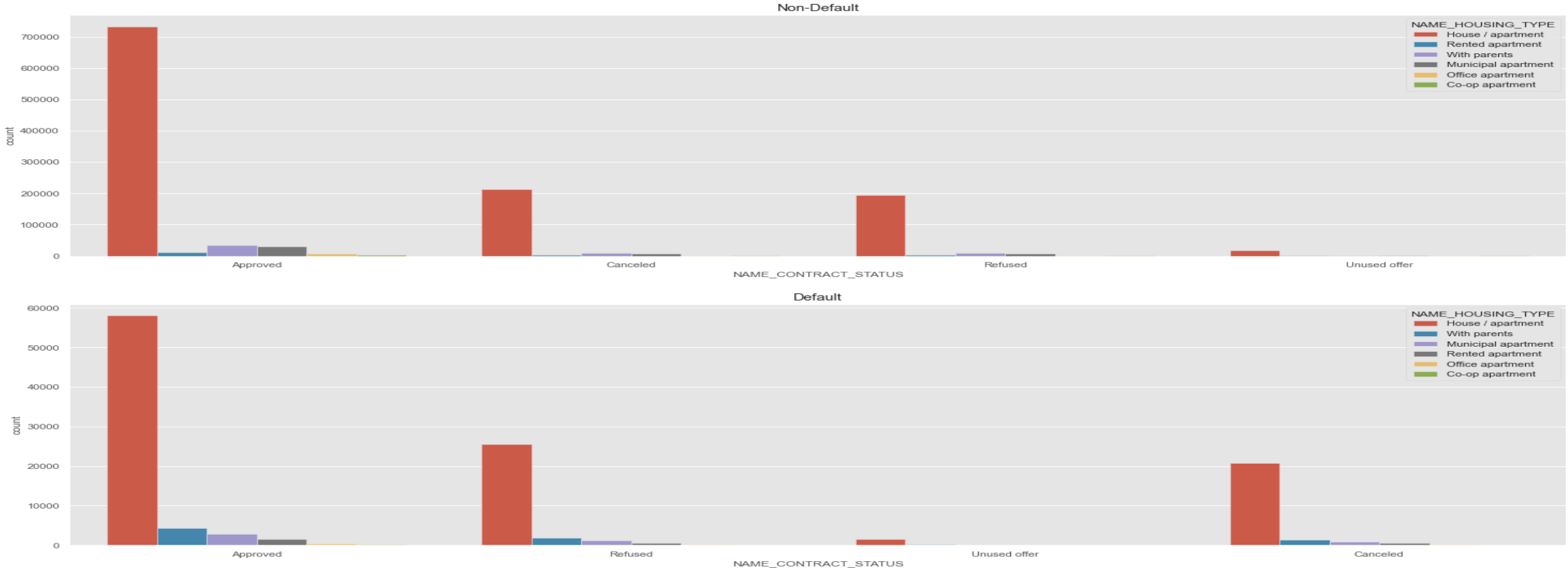
- Most of the Approved & Non-Defaulted applicants are Females & least are Males.
- Most of the Approved & Defaulted applicants are Females & least are Males.





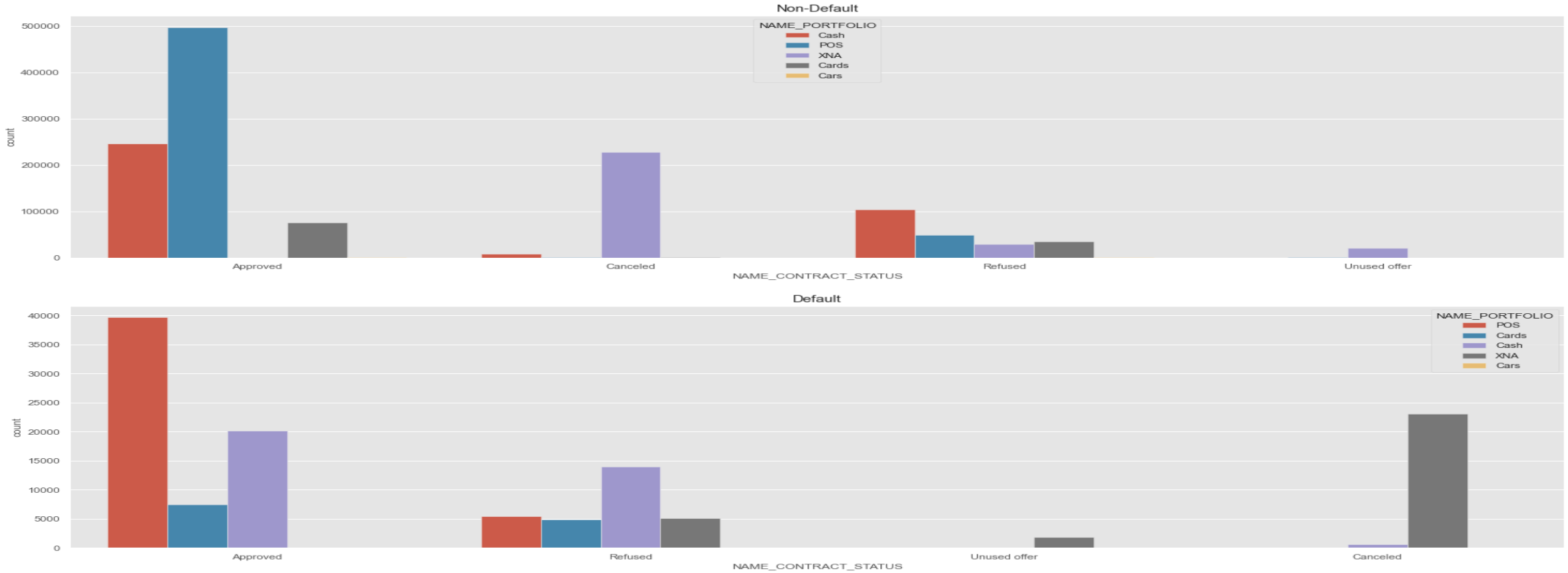
Name contract status vs Name Housing Type

- Most of the Non-Defaulted Credit is Approved for those who own a House/Apartment & the least approved those who stay at Co-op apartments.
- Most of the Defaulted Credit is Approved for those who own a House/Apartment & the least approved those who stay at Co-op apartments.



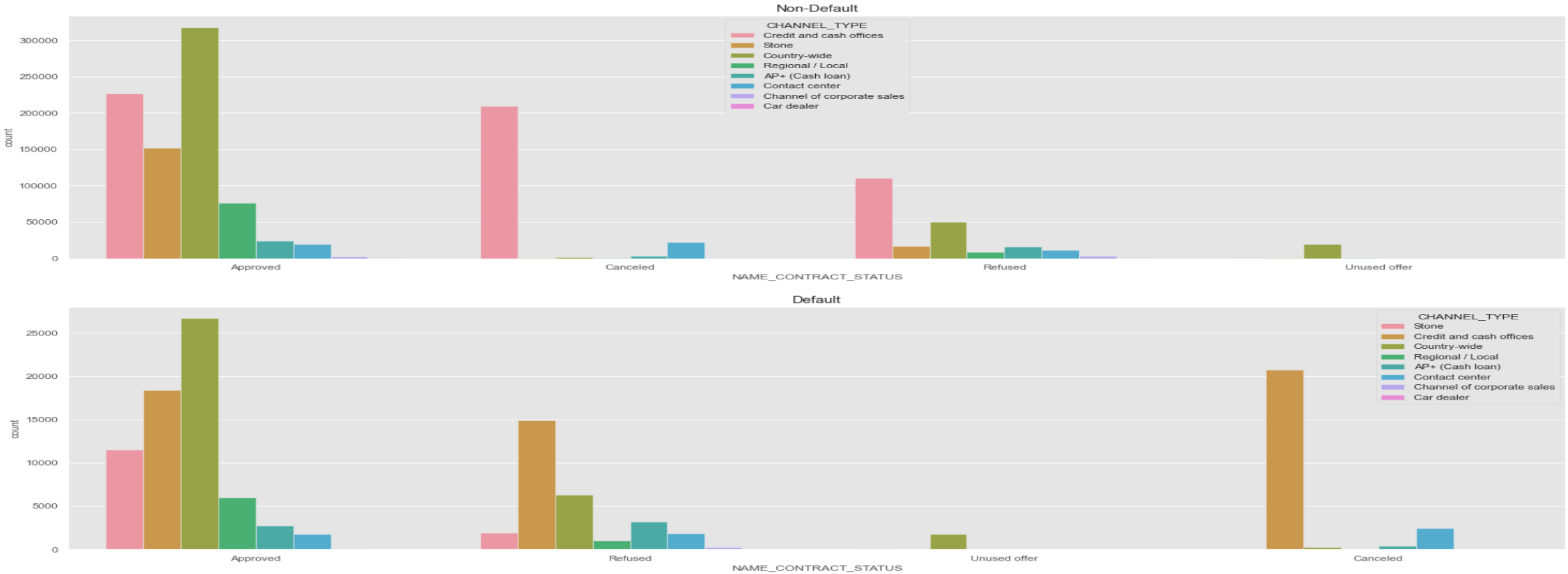
Name Contract Status vs Name Portfolio

- Most of the Non-Defaulted Credit is Approved for POS Portfolio & the least approved is for Cash.
- Most of the Defaulted Credit is Approved for POS & the least is approved for Cards.



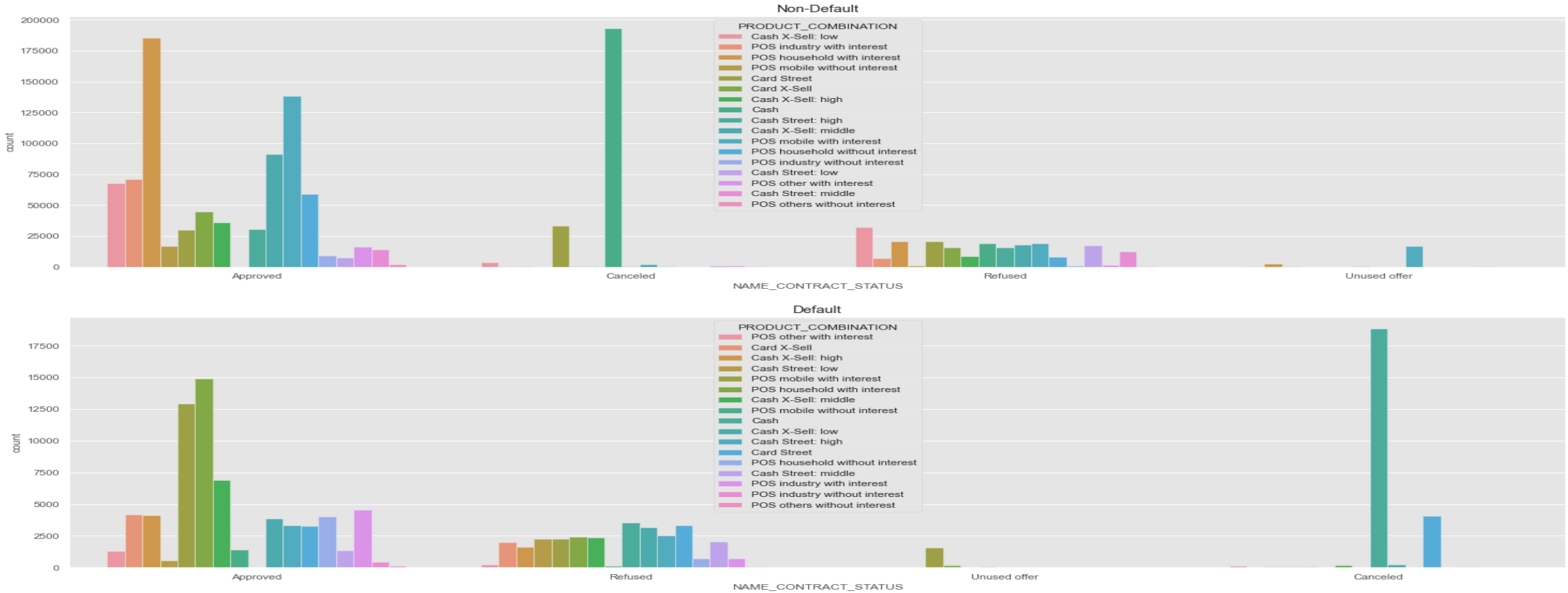
Name Contract status vs Channel Type

- Most of the Non-Defaulted Credit is approved in Previous application through the Channel - 'Country-Wide' & least is approve through "Channel of corporate sales".
- Most of the Defaulted Credit is approved through Country-Wide Channel & least is approved through Contact Number.

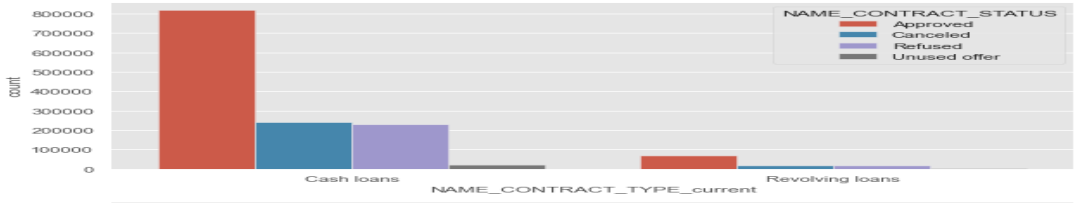


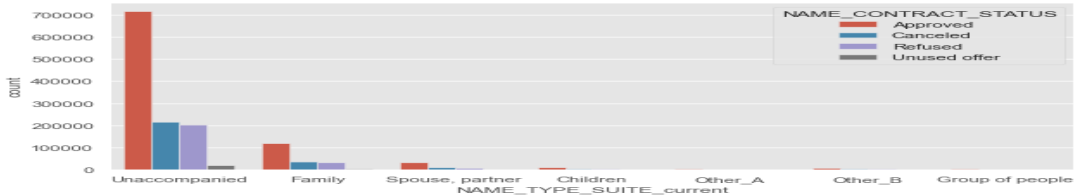
Name Contract Status vs Product combination

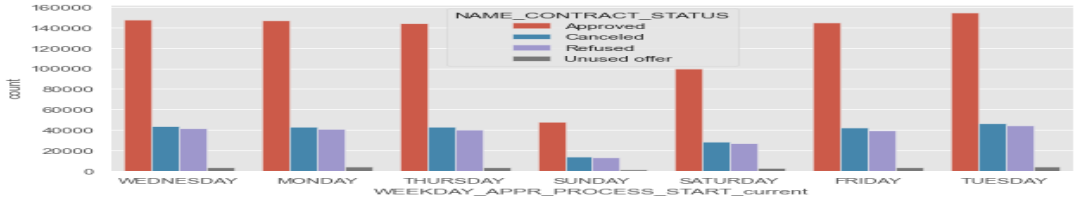
- Most of Non Defaulters Credit is Approved for Product is "POS household with interest" and the least approved Credit Product is "Cash Street Middle".
- Least Non-Default Credit denied/refused product is "POS other without Interest".
- Most of the Defaulted Credit is approved for "POS household with interest" and the least approved for Credit product is "POS others without interest".
- The least credit refused/canceled/unused Default product is Refused - 'Card Street".

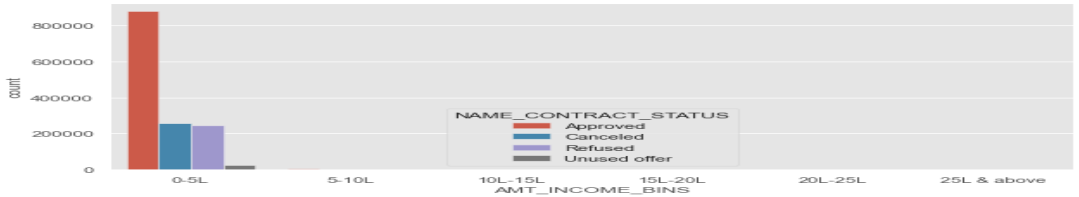


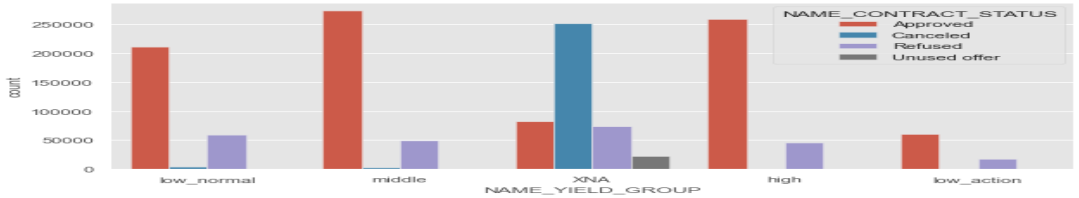
Categorical - Categorical with respective to Name Contract status

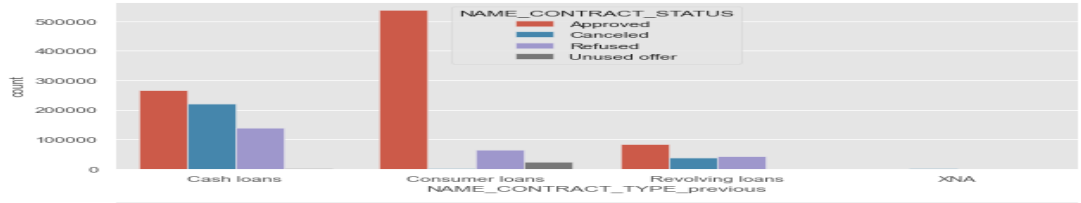


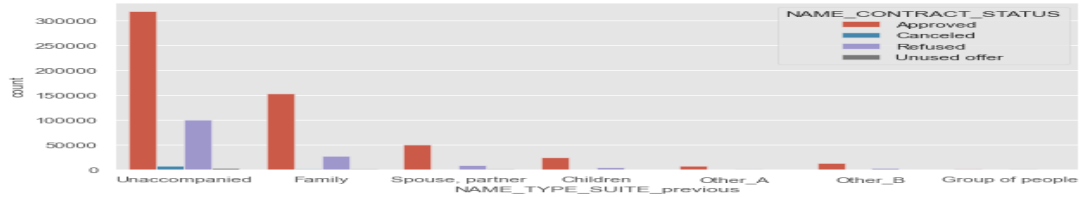


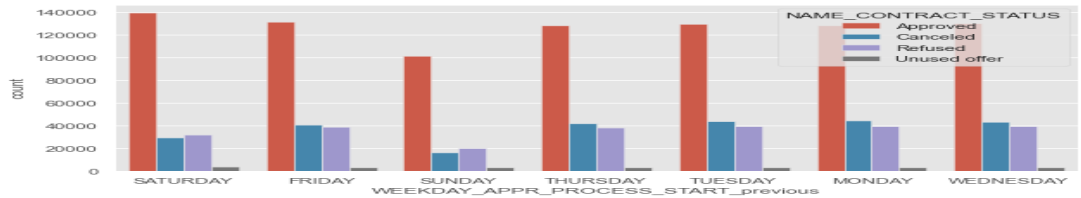


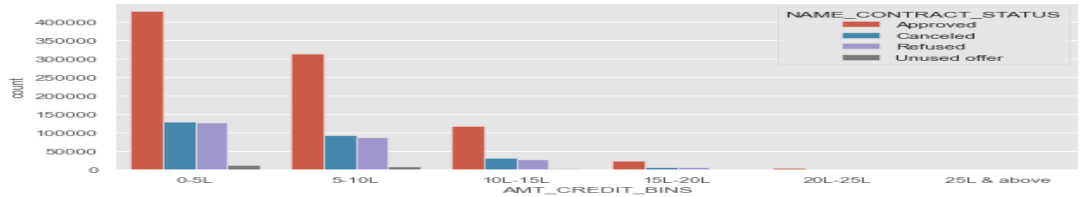


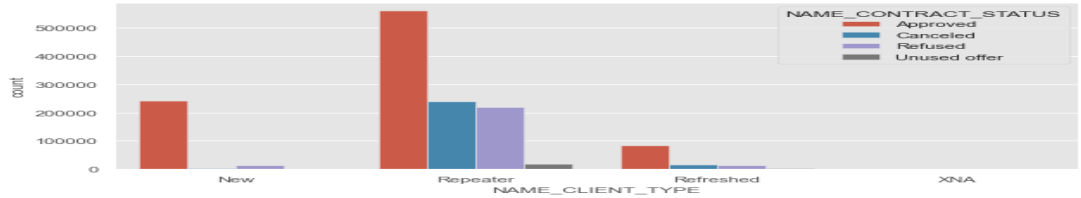










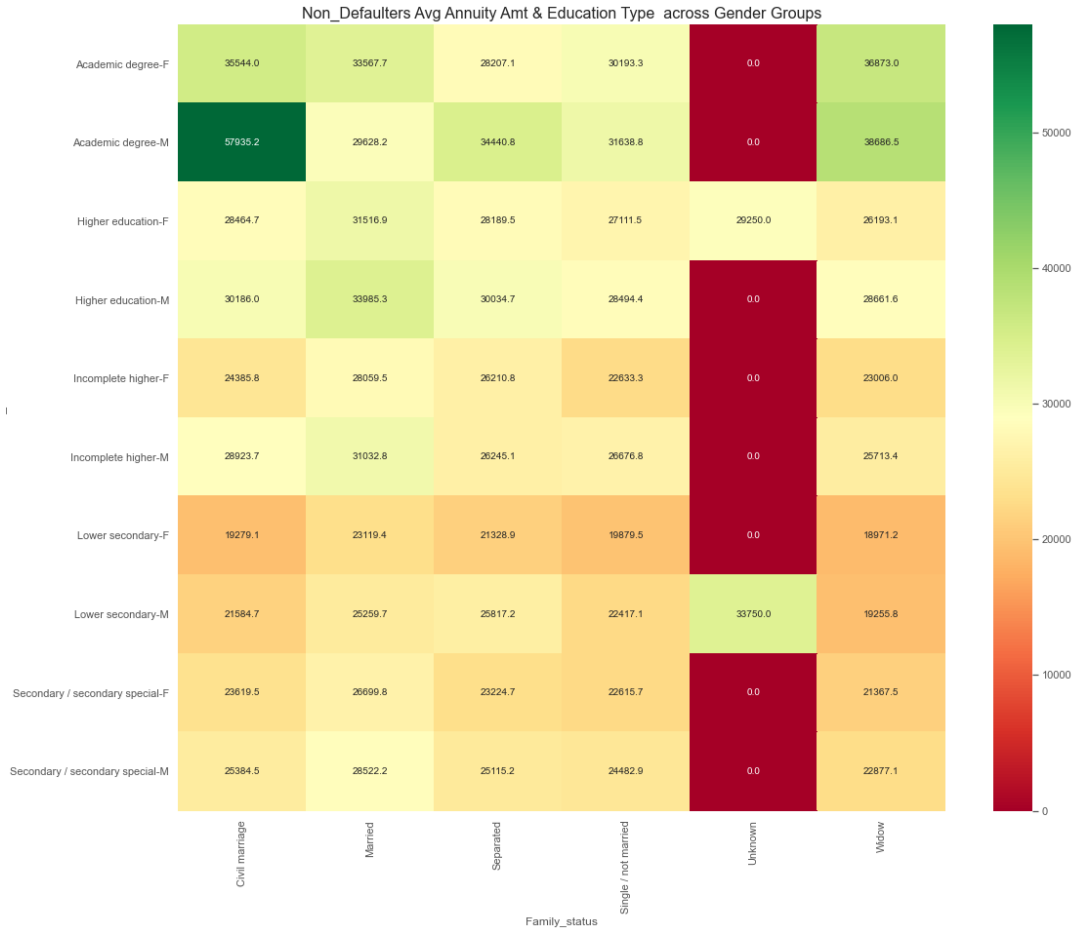


Categorical - Categorical with respective to Name Contract status(contd..)

- Repeater has highest number of approved loans.
- Middle NAME_YIELD_GROUP has highest approval.
- Value of AMT_CREDIT_BIN does not affect loan approvals.
- For Medium AMT-INCOME-TOTAL-bin the approval is highest in previous application Saturday has the highest approval rate, but in current application it is Tuesday.
- both in NAME CONTRACT TYPE Previous and NAME CONTRACT TYPE Current unaccompanied has the highest number.
- currently bank is only giving two types of loans -Cash and Revolving Loans.
- Previously bank was providing Cash, Revolving and Consumer loans.
- Number of consumer loans were highest previously and now highest number is Cash loans.

Multivariate Analysis (Across Multiple Numerical & Categorical Variables)

Avg Annuity Amt & Education Type across Gender Groups



Avg Annuity Amt & Education Type across Gender Groups (Excluding XNA)

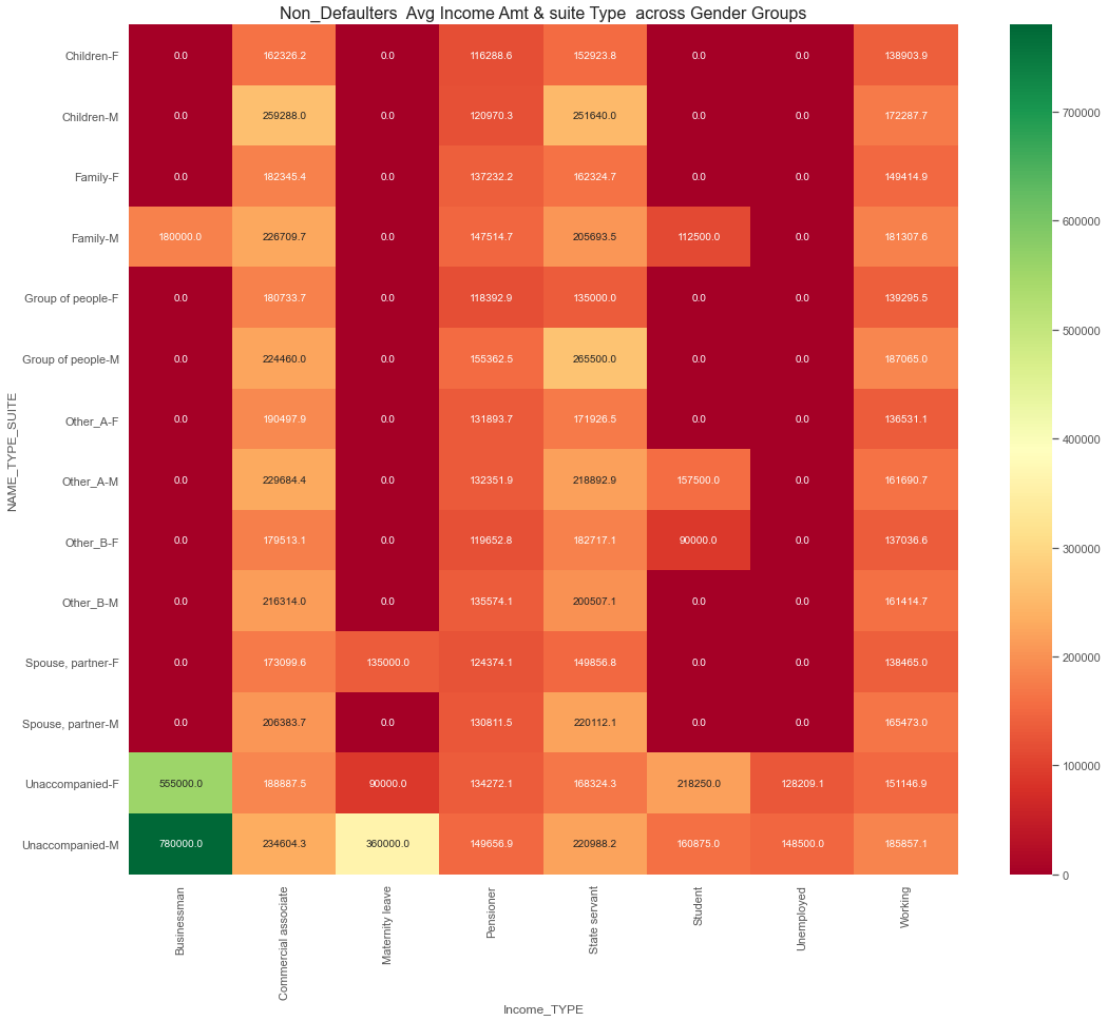
Non-Defaulters:

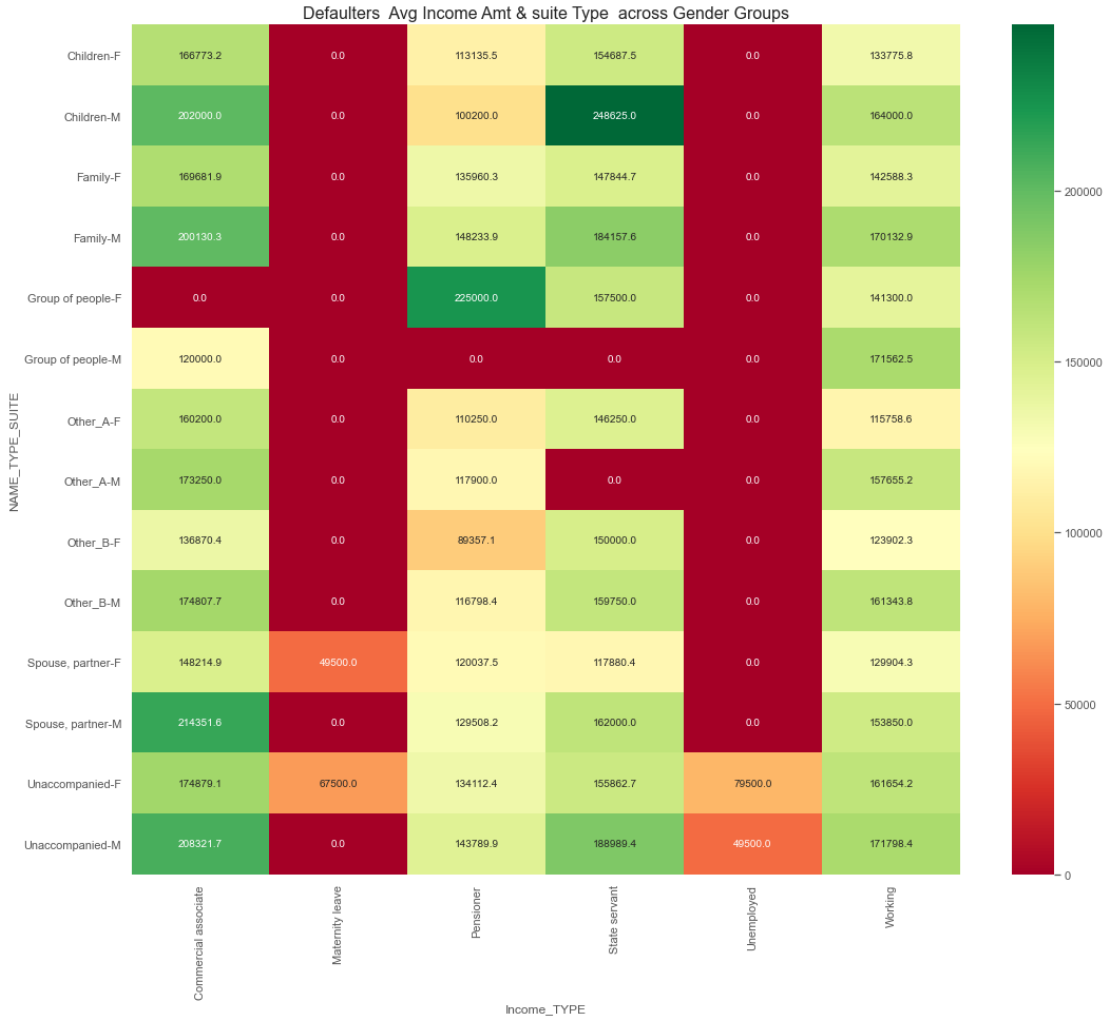
- Males with Academic Degree & Civil Marriage status paid the highest Annuity average across the Gender, Education types & Gender groups .
- Males with Academic Degree & Family Status - Widow & separated status group paid the next highest Annuity.
- Least amount Average Annuity is paid by Married clients with Secondary /Secondary Special Education Type.

Defaulters:

- Highest amount of Average Annuity is defaulted by Females with Academic Degree & Single/Not Married Status.
- Next Highest Average Annuity is defaulted by Males with Higher Education & Married status.
- Least average Annuity is defaulted by Females Lower/Secondary & Single /Not Married Status.

Average Income Amt & Suite Type across Gender Groups





Average Income Amt & suite Type across Gender Groups

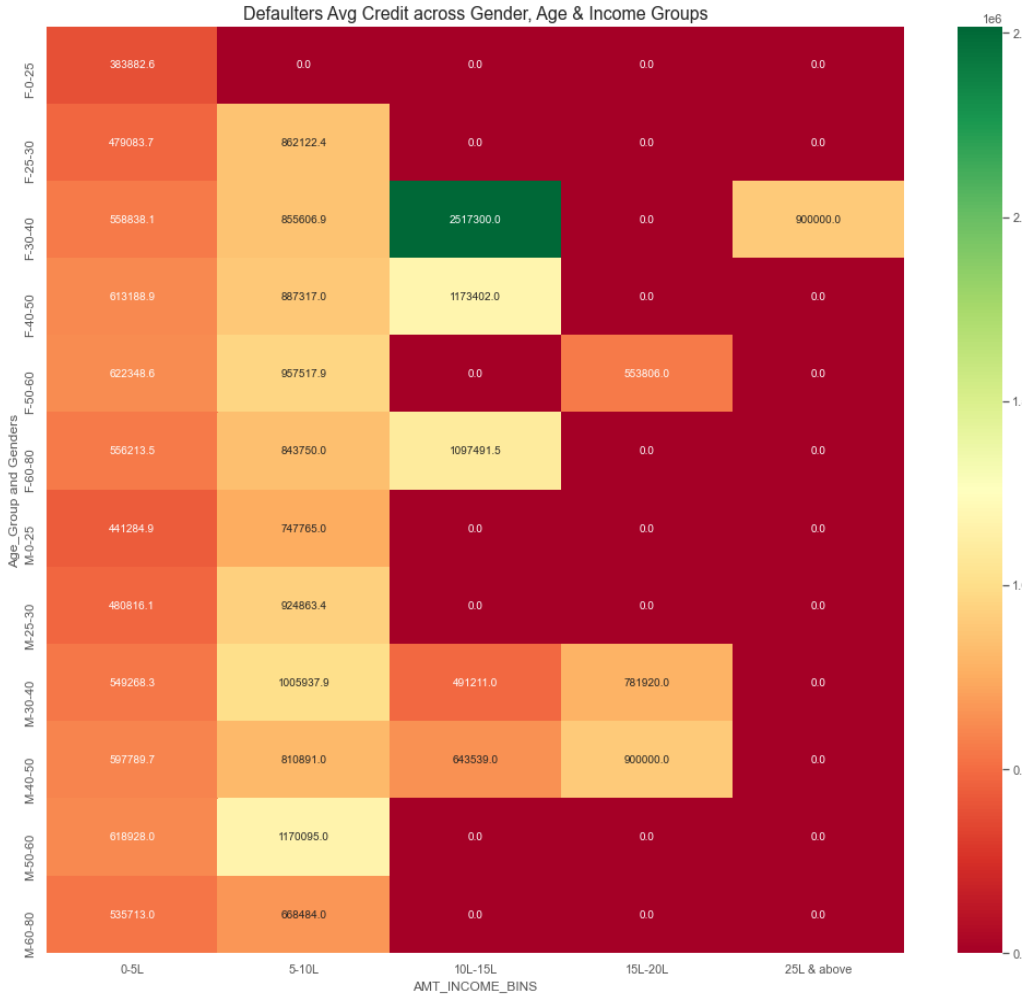
Non-Defaulters:

- Highest Avg Income groups among Non-Defaulters are Males, Businessman & Unaccompanied followed by Female, Businessman & Unaccompanied.
- All the other combinations appear to be the least Avg Income groups.

Defaulters:

- Highest Avg Income groups who defaulted the Credit are Male, State Servant with Children.
- The next highest Income group who defaulted is Female, Pensioners with Group of People.
- Its observed that all other good average Income groups like Commercial Associate, State Servant & Working groups across all the Accompanied groups also defaulted the Credit.
- The least average Income groups who defaulted are Females with Maternity Leave accompanied by Spouse/ Partner or Unaccompanied.

Average Credit across Age ,Gender& Income Groups



Average Credit across Age, Gender& Income Groups

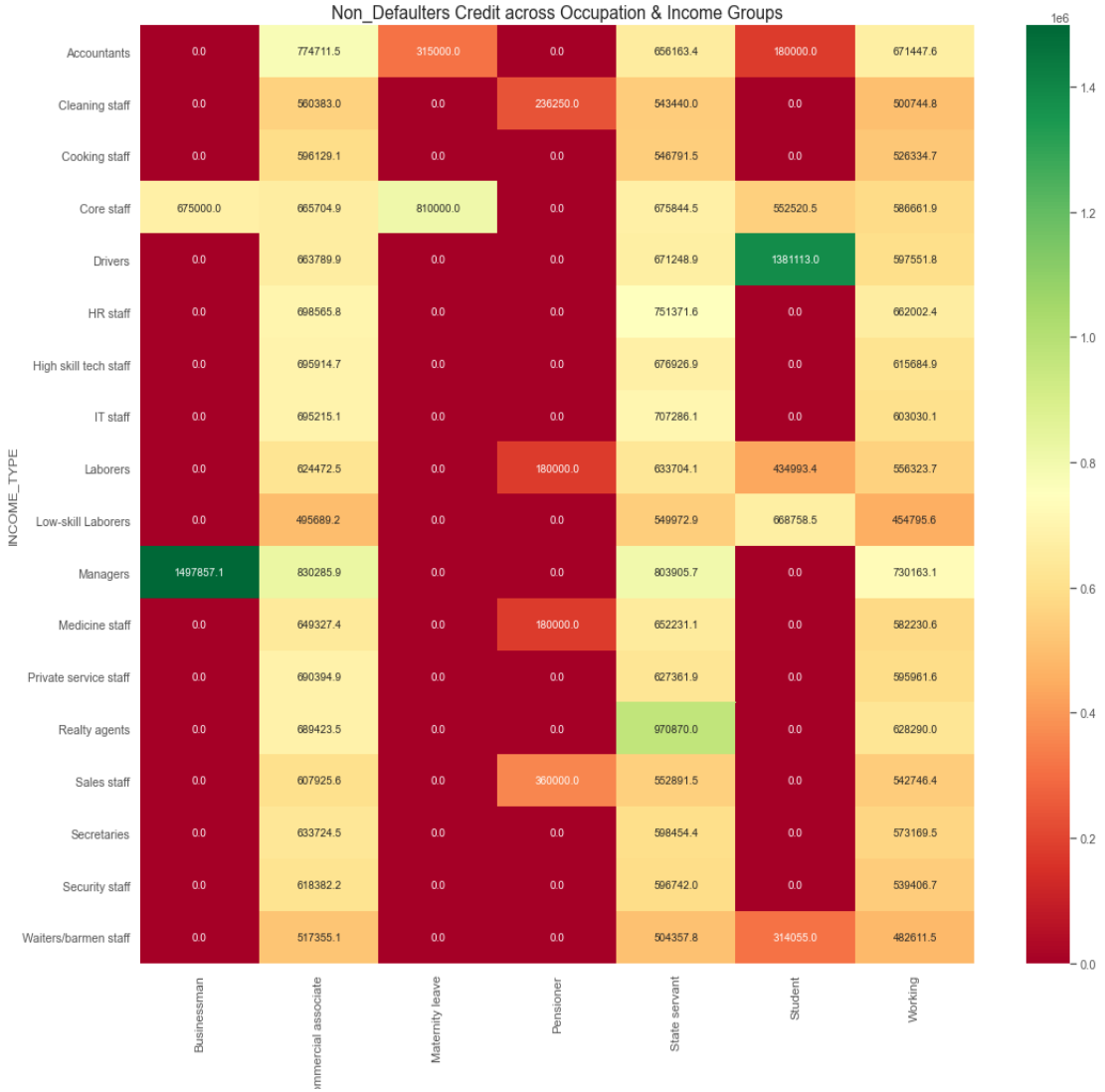
Non-Defaulters:

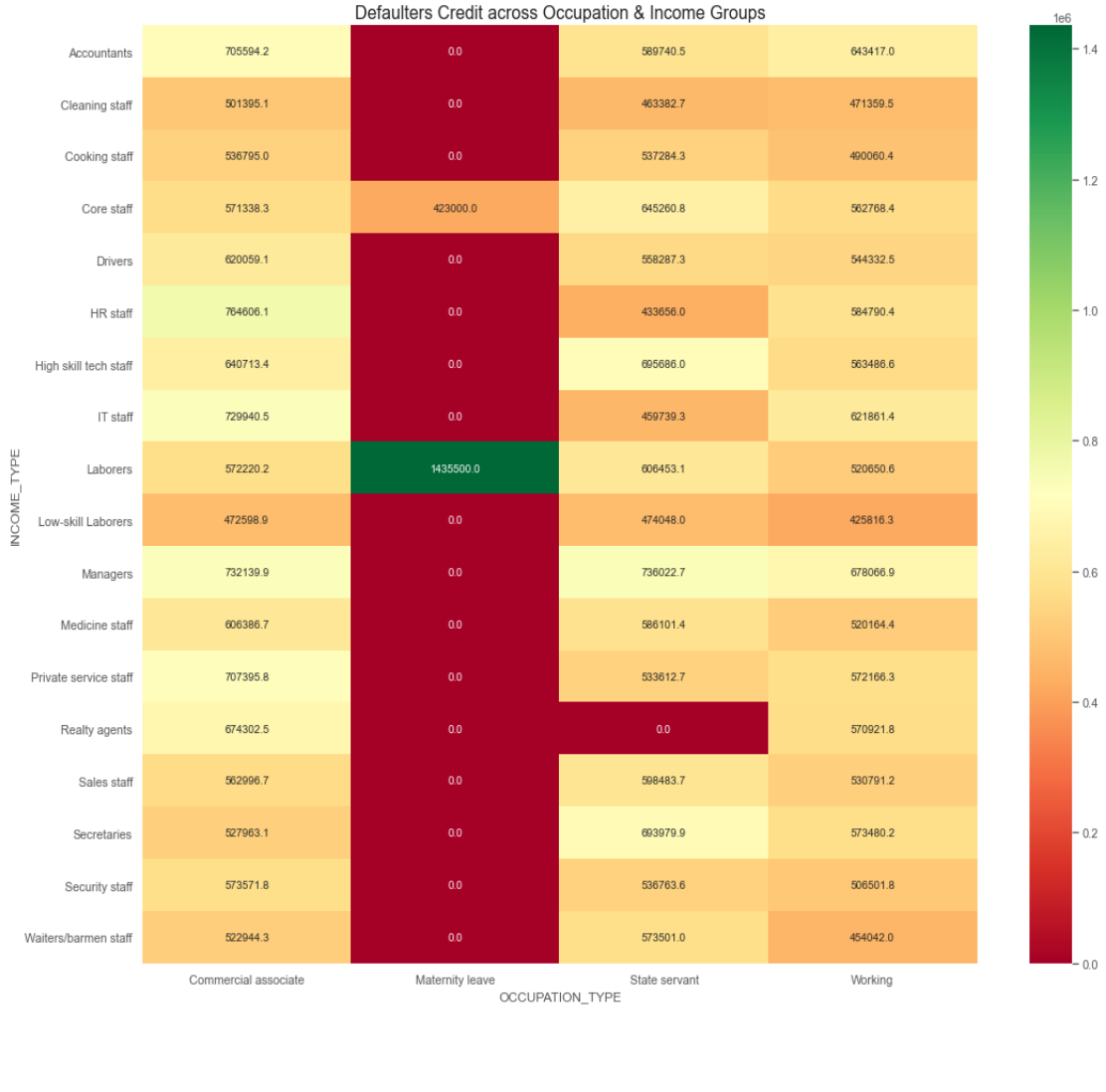
- Males with 40-50 age group & Income of 15-20 Lakhs have been approved the highest average Credit amount.
- Most of the Males & Females in 25-80 Range age groups & Income in 5-10Lakhs, 10-25Lakhs, 25Lakhs & Above groups were approved with maximum Credits.
- Least average Credit amount was approved for 0-5 Lakhs income groups irrespective of age & Gender.

Defaulters:

- Most of the defaulters approved with highest average Credit is Females in 30-40 age group with 10-15 Lakhs income.
- Least average Credit was approved for 0-10 Lakhs income range irrespective of age & gender.

Average Credit across Occupation Type& Income Type





Average Credit across Occupation Type& Income Type

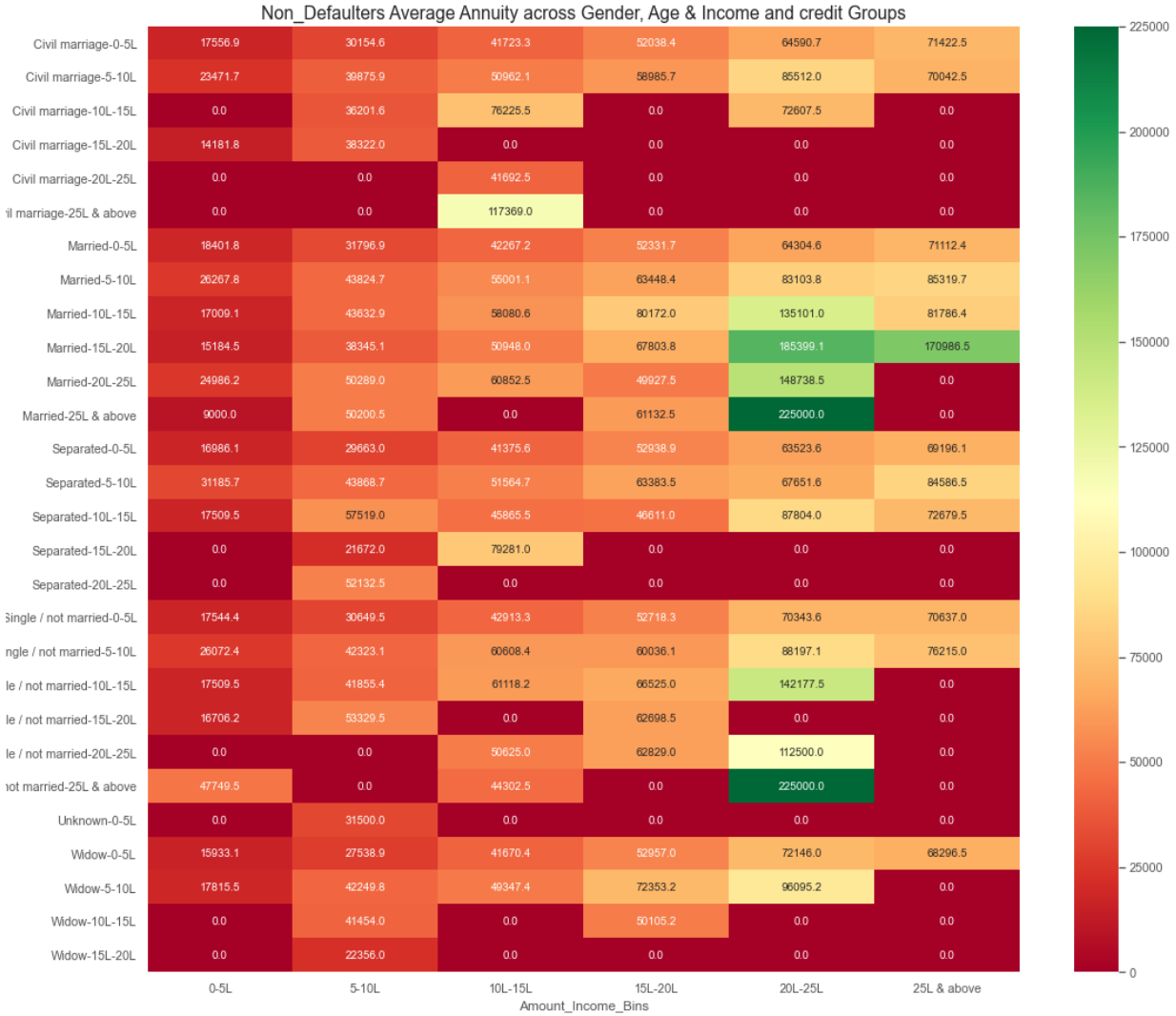
Non-Defaulters:

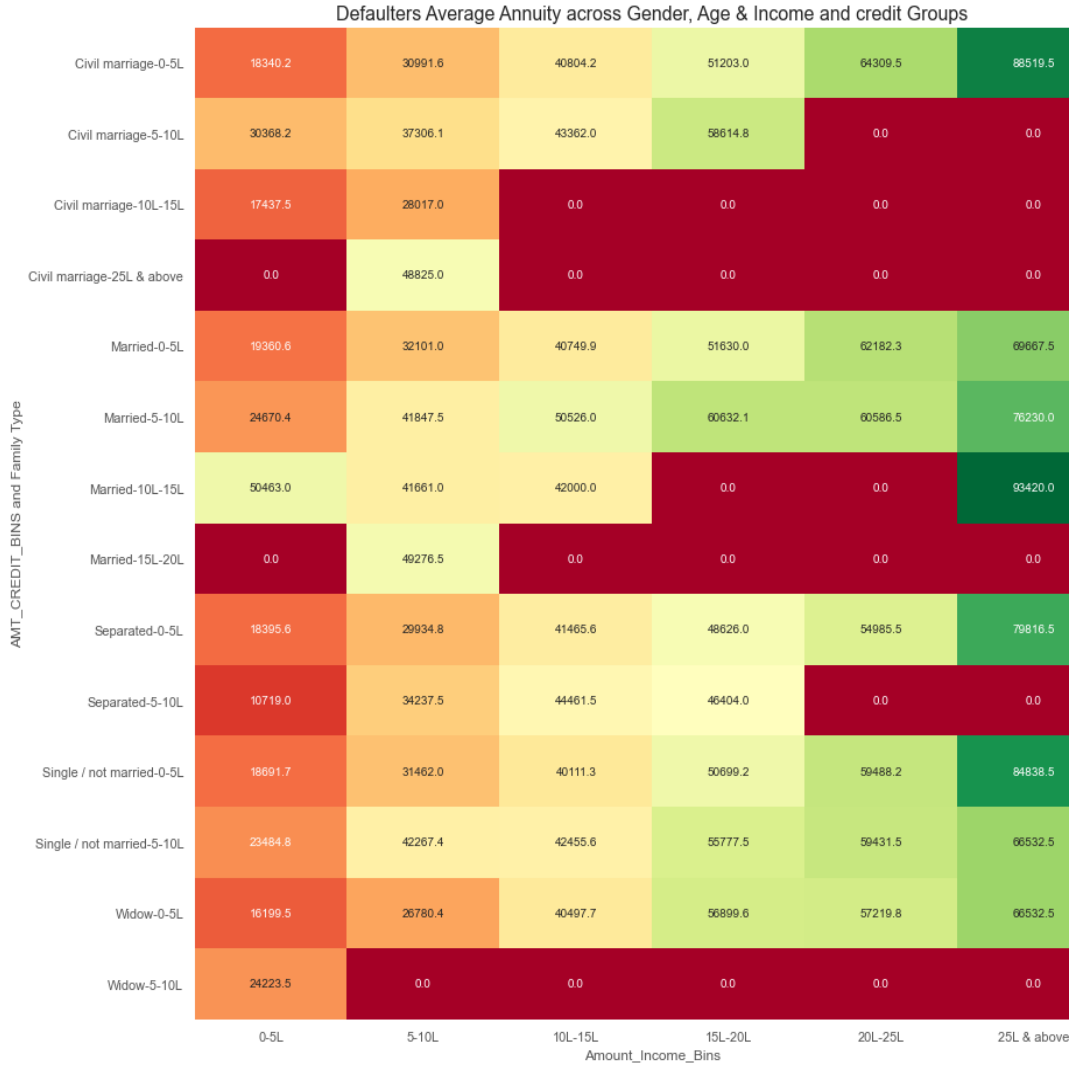
- Most of the Credit approved Non-Defaulters are Businessman by occupation & Managers by Income Type.
- Also Students by Occupation & Drivers by Income type have also taken Maximum average Credit.
- All other combinations irrespective of Occupation type & Income type were approved least Credits.

Defaulters:

- Maximum Credit Defaulted by Laborers on Maternity Leave.
- All other combinations of Occupation & Income Type are approved least credit amounts.

Average Annuity across Family-Status ,Income Range & Credit Range





Average Annuity across Family status , Income Range & Credit Range

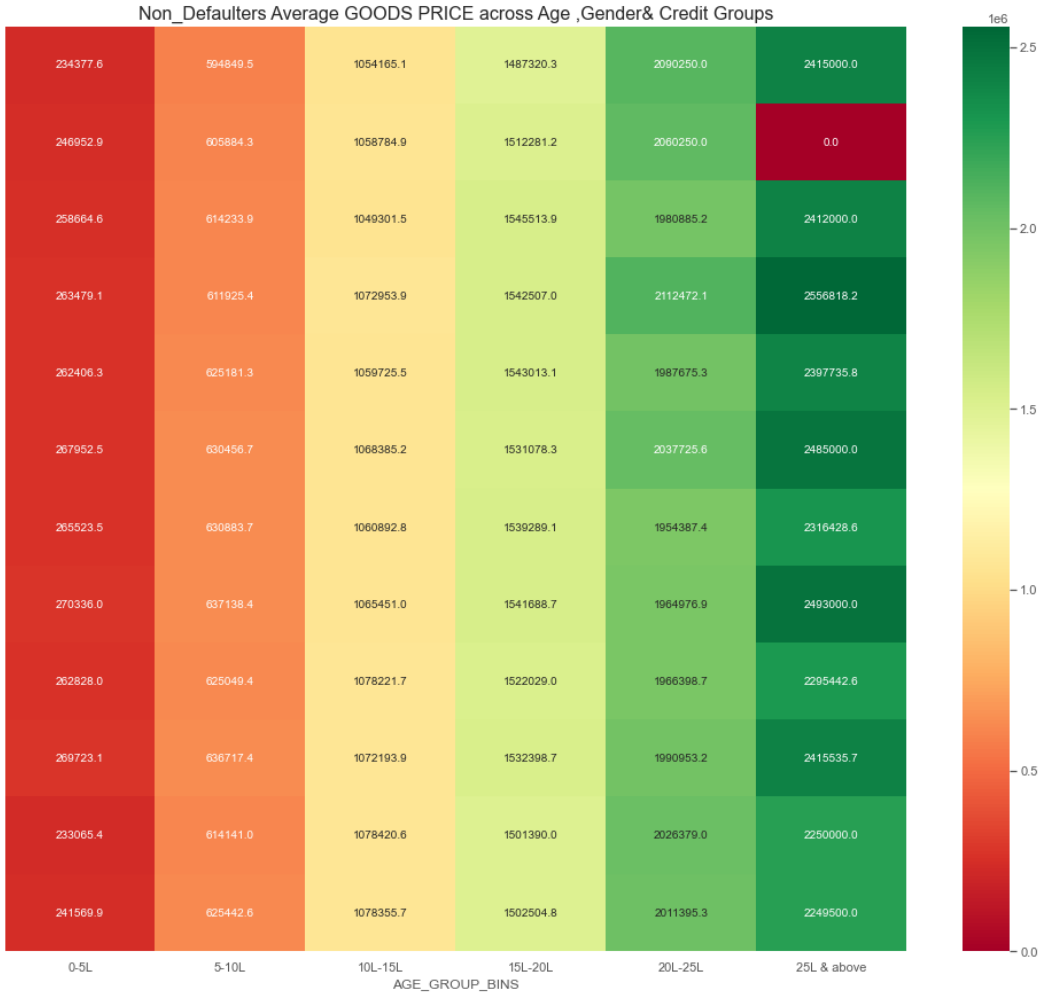
Non-Defaulters:

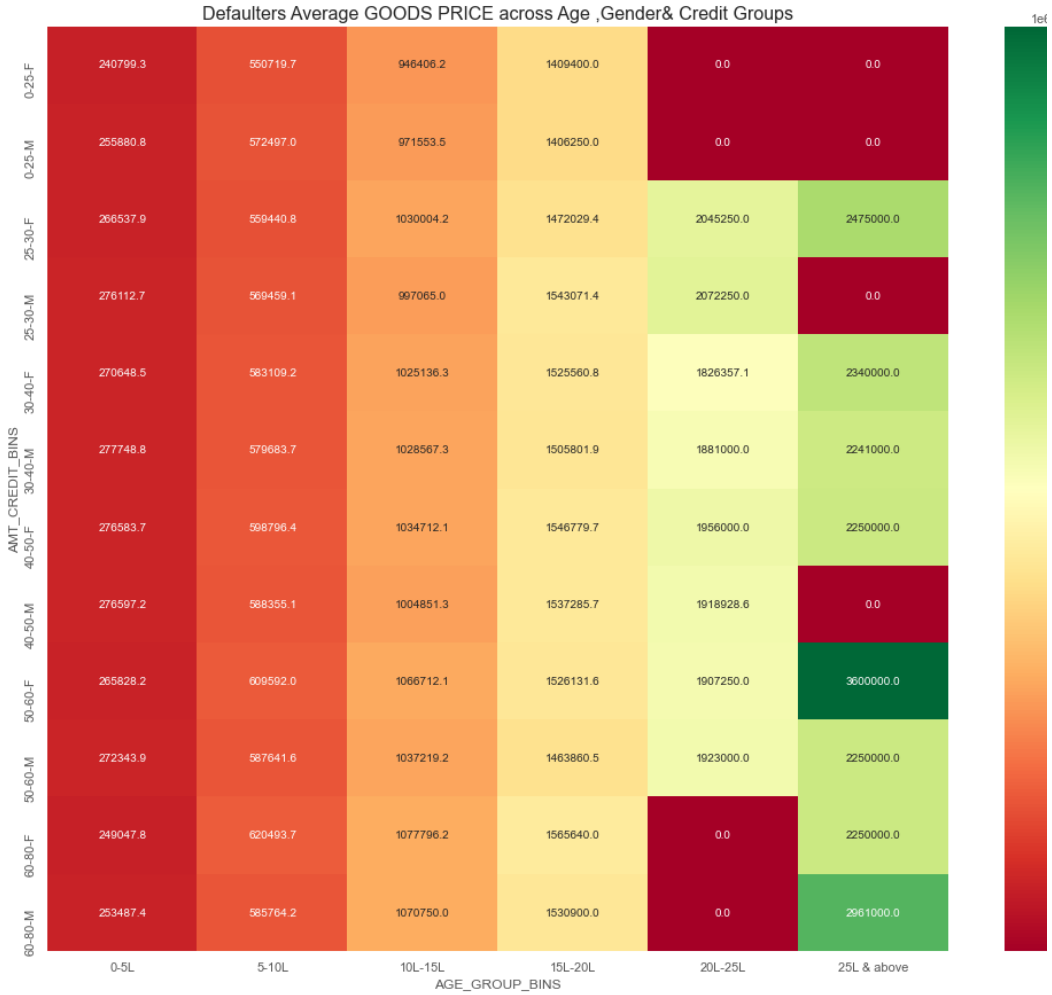
- Highest Average Annuity is paid by Single /Not Married with Income 20-25 Lakhs & Credit amount 25Lakhs & above group and Married with Income range 20-25 Lakhs and having Credit 25Lakhs & above.
- In 20-25 Lakhs Income Range, Single with Credit 10-15 Lakhs, Married with Credit 10-25 Lakhs also have paid Highest Annuity.
- Least Average Annuity is paid by all Family status Type groups irrespective of the Income & Credit amounts.

Defaulters:

- Defaulters who have defaulted the highest Average Annuity with Income 25Lakhs & above, Married with Credit of 10-15Lakhs and Civil Marriage with 0-5 Lakhs Credit Range groups.
- Defaulters who also defaulted good Average Annuity are in 10-25 Lakhs Income Group & 0-25Lakhs Credit group.
- The least average annuity Defaulters are with 0-10 Lakhs Income range & with all types of family status & credit ranges.

Average GOODS PRICE across Age ,Gender& Credit Groups





Average GOODS PRICE across Age ,Gender& Credit Groups

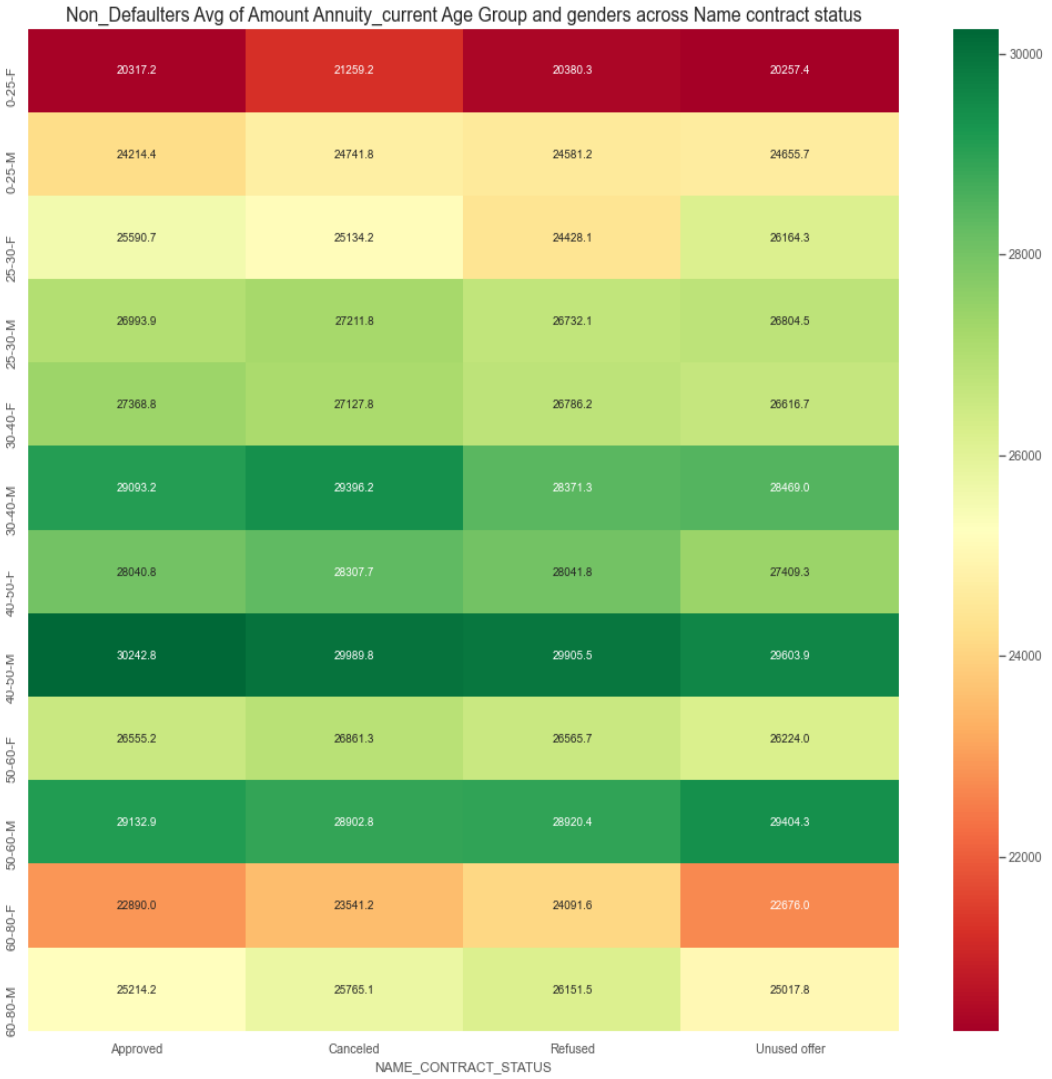
Non-Defaulters:

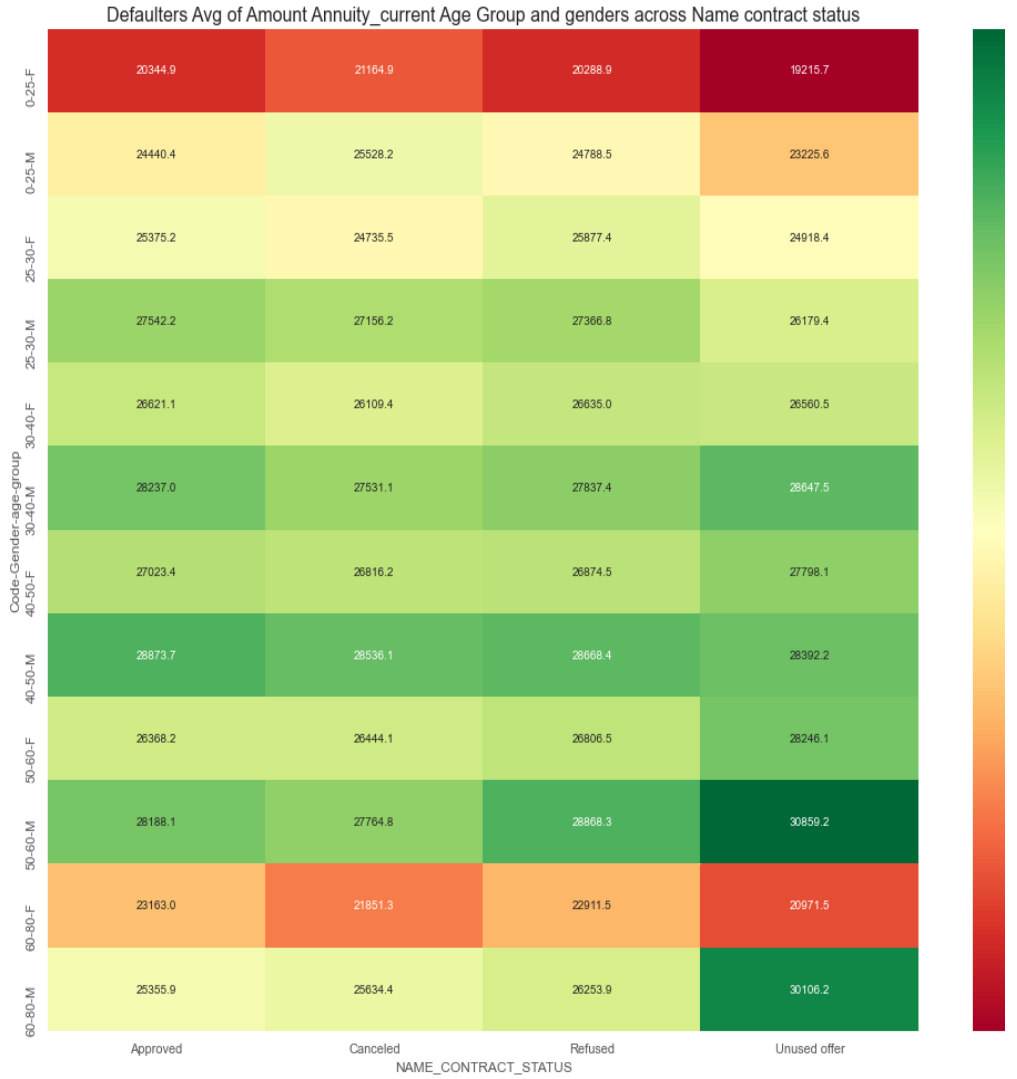
- Most of the Males in 20-30, 30-40 ,40-50 & 50-60 age groups with Credit of 25Lakhs & above have applied for the Maximum Average of the AMT Goods Price.
- Most of the Non-Defaulters with Credit above 20Lakhs & irrespective of gender & all age groups except for 0-25 Males have applied for Average Goods Price.
- 0-5Lakhs income groups have taken the least average Goods Price loans.

Defaulters:

- Most of the Defaulter who have taken the maximum average goods price loan is Females,50-60 age with Credit of 25Lakhs & above.
- Followed by most of the other Defaulters is Males in 30-40 years ,50-60 years & 60-80 years range with credit of 25Lakhs & above.
- Least Average Goods Price is taken by 0-5 Lakhs Credit range irrespective of the Gender & Age groups.

Average Annuity current across Age, Gender & Contract Status





Average Annuity current Age Group and genders across Name contract status

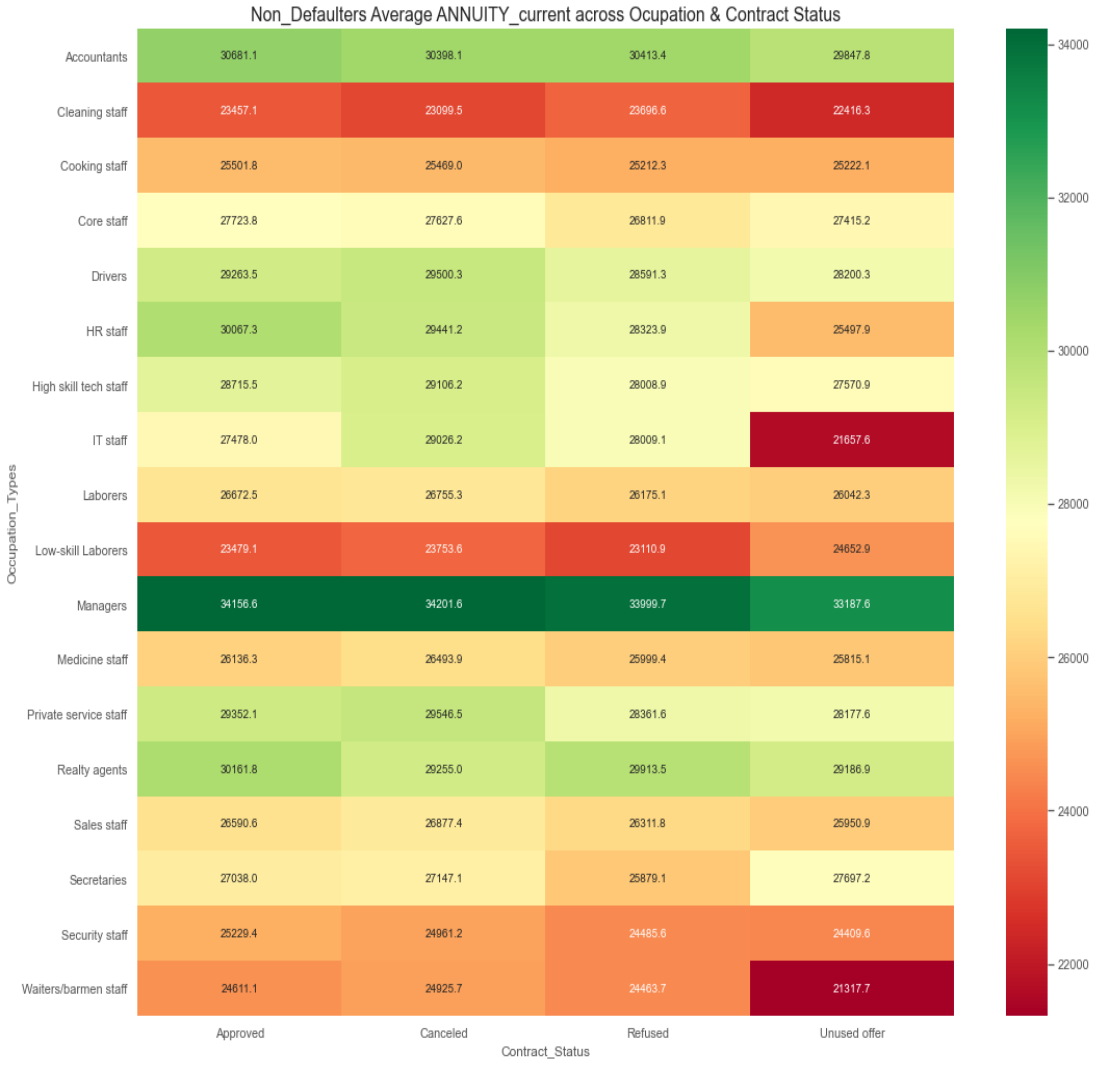
Non-Defaulters:

- Most of the Non-Defaulters who paid the highest Annuity current is Males in 40-50 age group. Followed by 30-40 age group Males, 40-50 age group Females & 50-60 age group Males for all the application status.
- Least Amount of Annuity current is paid by 0-25 age group females in all application status.

Defaulters:

- Most of Defaulters who defaulted the highest Annuity current amount is 50-60 age group Males with the "Unused Offer" status.
- All other groups who also defaulted good average annuity current is 25-60 age group Males & Females irrespective of application status.
- Least average Annuity current amount is defaulted by 0-25 age group Females and 60-80 age group Females irrespective of application status.

Average ANNUITY Current across Occupation & Contract Status



Average ANNUITY current across Occupation & Contract Status

Non-Defaulters:

- Most of the Average Annuity current is paid by Managers irrespective of the previous application status.
- Least Average Annuity current is paid by Cleaning Staff & Low Skilled Laborers.

Defaulters:

- Most of the Average NetCurrent is defaulted by Reality Agents & Managers.
- Least of the Average Annuity current is defaulted by Cleaning Staff & Low Skilled Laborers.

Conclusion

- Majority of the Defaulters Credit amount is in 0-5 Lakhs range with 50.8% & least Credit Amount is 25Lakhs & above with 0.1%.
- Most of the Defaulters in 30-40 age group Opted Cash Loans for Goods Price Loans with Annuity between 2-2.5 Lakhs
- Among Defaulters, Businessman pay the highest Annuity & Students pay the least Annuity.
- Majority of the Defaulters with Higher Education and defaulted an Annuity between 2-2.5 Lakhs
- Most of the Defaulters who own House/Apartment got the Credit approved and don't have a car.
- Most of the Defaulters are Married & Laborers by Occupation and least are Widow & Laborers.

Conclusion(contd..)

- Highest amount of Average Annuity is defaulted by Females with Academic Degree & Single/Not Married Status
- Highest Avg Income groups who defaulted the Credit are Male, State Servant with Children
- Most of the defaulters approved with highest average Credit is Females in 30-40 age group with 10-15 Lakhs income
- Maximum Credit Defaulted by Laborers on Maternity Leave.
- Defaulters who have defaulted the highest Average Annuity with Income 25Lakhs & above are Married with Credit of 10-15Lakhs and Civil Marriage with 0-5 Lakhs Credit Range groups.
- Most of the Defaulter who have taken the maximum average goods price loan is Females,50-60 age with Credit of 25Lakhs & above.

Final Words

Target/focused variable for Application dataset - **TARGET**
Target/focused variable for Previous dataset **- NAME_CONTRACT_STATUS**

Top Major variables to consider for loan prediction are :

- NAME_EDUCATION_TYPE
- NAME_INCOME_TYPE
- NAME_HOUSING_TYPE
- NAME_GOODS_TYPE
- CNT_CHILDREN
- CNT_FAM_MEMBERS
- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE
- DAYS_BIRTH
- DAYS_EMPLOYED
- DAYS_REGISTRATION

Final Words (Contd..)

In future we can build a Recommendation Engine using Machine Learning which provides if the Customers are Defaulters or Non-Defaulters.

- For Recommendation Engine, we can consider the NAME_EDUCATION_TYPE ,NAME_INCOME_TYPE, NAME_HOUSING_TYPE ,NAME_GOODS_ TYPE, CNT_CHILDREN , CNT_FAM_MEMBERS , AMT_INCOME_TOTAL , AMT_CREDIT , AMT_ANNUITY , AMT_GOODS_PRICE , DAYS_BIRTH , DAYS_EMPLOYED & DAYS_REGISTRATION.
- User will enter the Customer Details in User Interface then, it will run the Recommendation Engine / Machine Learning Algorithm in the backend and it displays if the customer is eligible for the Credit or not. This Web Application will be helpful for the Bank Mangers to predict and it will reduce the Risk.

Thank You