Credit Risk Analysis

(Group Case Study)

By

Akansha Khandelwal

And

Neelanjan Basu

04/08/2021

Agenda

Business Objective

Summary

- The EDA analyze the current and historical loan data, to let the loan providing companies (or banks) know the driving parameters/indicators based on which the bank employees can decide on the loan application based on the applicant's profile and credit history (if any).

- There are mainly two risks involved with the bank decision
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
  - 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.

Dataset and its Parameters

As part of the case study 3 datasets were provided as follows

- *'application_data.csv'* contains all the information of the client at the time of application. The data is about whether a **client has payment difficulties.**
- *'previous_application.csv'* contains information about the client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**
- *'columns_description.csv'* is data dictionary which describes the meaning of the variables.

5 Dataset and its Parameters (continued…)

As part of the analysis, ***'application_data.csv'*** contains around 122 columns and ***'previous_application.csv'*** contains around 37 columns, which we would use here as part of our analysis.

- For the dataset ***'application_data.csv'*** there are around **49 columns** which have more than 45% null values
- For the dataset ***'previous_application.csv'*** there are around **05 columns** which have more than 40% null values
- *As part of the data understanding and cleaning, we will drop columns having nulls values > 45% from* ***'application_data.csv'*** *and > 40% for* ***'previous_application.csv'*** *while merging .*

COMMONAREA_AVG	69.872297
COMMONAREA_MEDI	69.872297
COMMONAREA_MODE	69.872297
NONLIVINGAPARTMENTS_MEDI	69.432963
NONLIVINGAPARTMENTS_MODE	69.432963
NONLIVINGAPARTMENTS_AVG	69.432963
FONDKAPREMONT_MODE	68.386172
LIVINGAPARTMENTS_MEDI	68.354953
LIVINGAPARTMENTS_AVG	68.354953
LIVINGAPARTMENTS_MODE	68.354953
FLOORSMIN_MODE	67.848630
FLOORSMIN_MEDI	67.848630
FLOORSMIN_AVG	67.848630
YEARS_BUILD_MODE	66.497784
YEARS_BUILD_AVG	66.497784
YEARS_BUILD_MEDI	66.497784
OWN_CAR_AGE	65.990810
LANDAREA_MODE	59.376738
LANDAREA_MEDI	59.376738
LANDAREA_AVG	59.376738
BASEMENTAREA_MODE	58.515956
BASEMENTAREA_MEDI	58.515956
BASEMENTAREA_AVG	58.515956
EXT_SOURCE_1	56.381073
NONLIVINGAREA_MODE	55.179164
NONLIVINGAREA_AVG	55.179164
NONLIVINGAREA_MEDI	55.179164
ELEVATORS_AVG	53.295980
ELEVATORS_MEDI	53.295980
ELEVATORS_MODE	53.295980
WALLSMATERIAL_MODE	50.840783
APARTMENTS_MEDI	50.749729
APARTMENTS_AVG	50.749729
APARTMENTS_MODE	50.749729
ENTRANCES_AVG	50.348768
ENTRANCES_MEDI	50.348768
ENTRANCES_MODE	50.348768
LIVINGAREA_MEDI	50.193326
LIVINGAREA_MODE	50.193326
LIVINGAREA_AVG	50.193326
HOUSETYPE_MODE	50.176091
FLOORSMAX_AVG	49.760822
FLOORSMAX_MEDI	49.760822
FLOORSMAX_MODE	49.760822
YEARS_BEGINEXPLUATATION_AVG	48.781019
YEARS_BEGINEXPLUATATION_MODE	48.781019
YEARS_BEGINEXPLUATATION_MEDI	48.781019
TOTALAREA_MODE	48.268517
EMERGENCYSTATE_MODE	47.398304
dtype: float64	

'application_data.csv' **null value columns > 45%**

RATE_INTEREST_PRIVILEGED	99.643698
RATE_INTEREST_PRIMARY	99.643698
RATE_DOWN_PAYMENT	53.636480
AMT_DOWN_PAYMENT	53.636480
NAME_TYPE_SUITE	49.119754

'previous_application.csv' **null value columns > 45%**

For our benefit during the EDA, we have categorized the individual dataset columns as below.

Dataset: ***'application_data.csv'***

- User Demographics
  - *Gender*
  - *House/Car Ownership*
  - *Age*
  - *Family*
  - *Education*
  - *Income*
  - *Employment*
  - *Marital Status*
  - *Residence*
  - *Contact Rating*
- User Actions / Ratings and Social Value
  - *ID change*
  - *External Rating*
  - *Social Rating/Observations*
  - *Loan Enquiries*
- Loan Summary
  - *Loan Type*
  - *Loan Amount*
  - *Loan Process*
  - *Loan Documents*

For our benefit during the EDA, we have categorized the individual dataset columns as below.

Dataset: ***'previous_application.csv'***

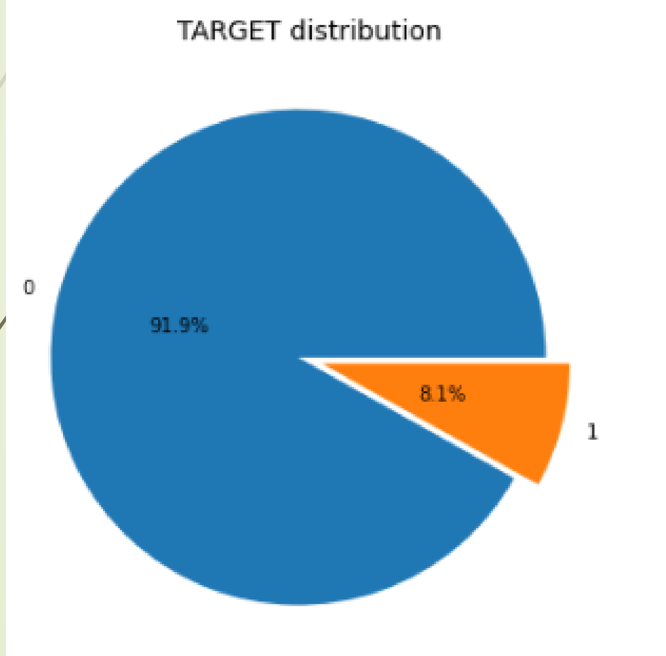
- Loan Summary
  - Loan Type
  - Loan Status
  - Loan Purpose
  - Loan Process
  - Loan Amount
  - Loan Interest
  - Loan Repayment

*Please note to keep the presentation short we will not add any univariate/bivariate analysis of the dataset **'previous_application.csv'**. All can be found in the 2nd Jupyter notebook.*

Anything variables required have fetched it in the main dataset and analyzed it at the end.

Total Applicant - Defaulter / Non-Defaulter Analysis

Defaulter / Non-Defaulter Percentage



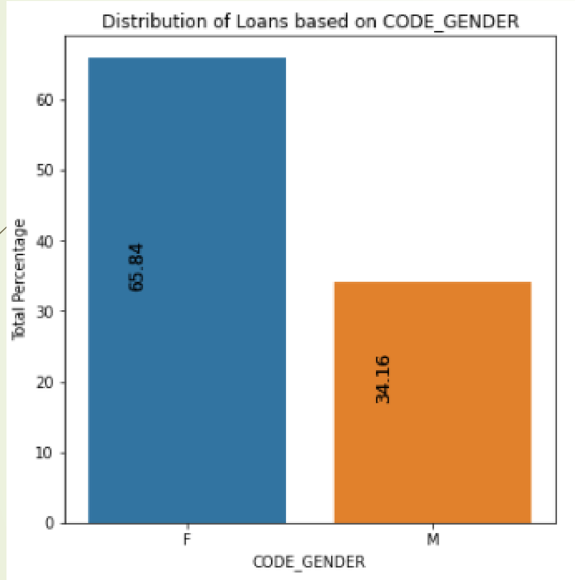
Loan Application % (Non-Defaulter)	Loan Application % (Defaulter)
91.9%	8.1%

Summary:

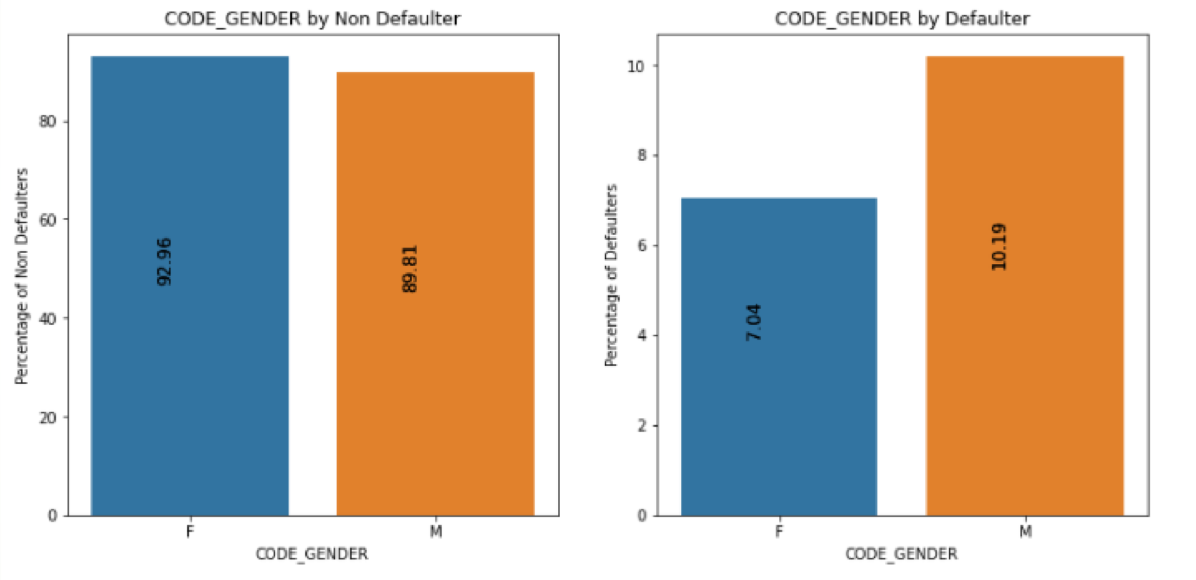
- Overall, 8.1% of the population have defaulted or having payment difficulties in repaying the loans

Gender Analysis - User Demographics

Gender % Vs Loan Applications



Gender % Vs Loan Applications (Defaulter/Non-Defaulter)



Gender	Loan Application %	Loan Application % (Non-Defaulter)	Loan Application % (Defaulter)
Male(M)	34.16%	89.81%	10.19%
Female(F)	65.84%	92.96%	7.04%

Gender Analysis - User Demographics (continued..)

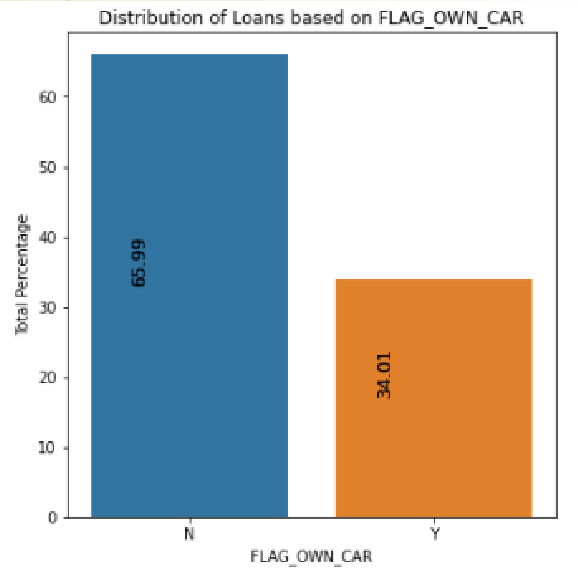
Summary:

- Gender contains 0% null values
- Females (65.75%) have more loan applications than Males (34.25%).
- Males (10%) tends to default loans more than Females(7.04%). Therefore, even Females apply for loans more than Males , Males tend to default more than Females

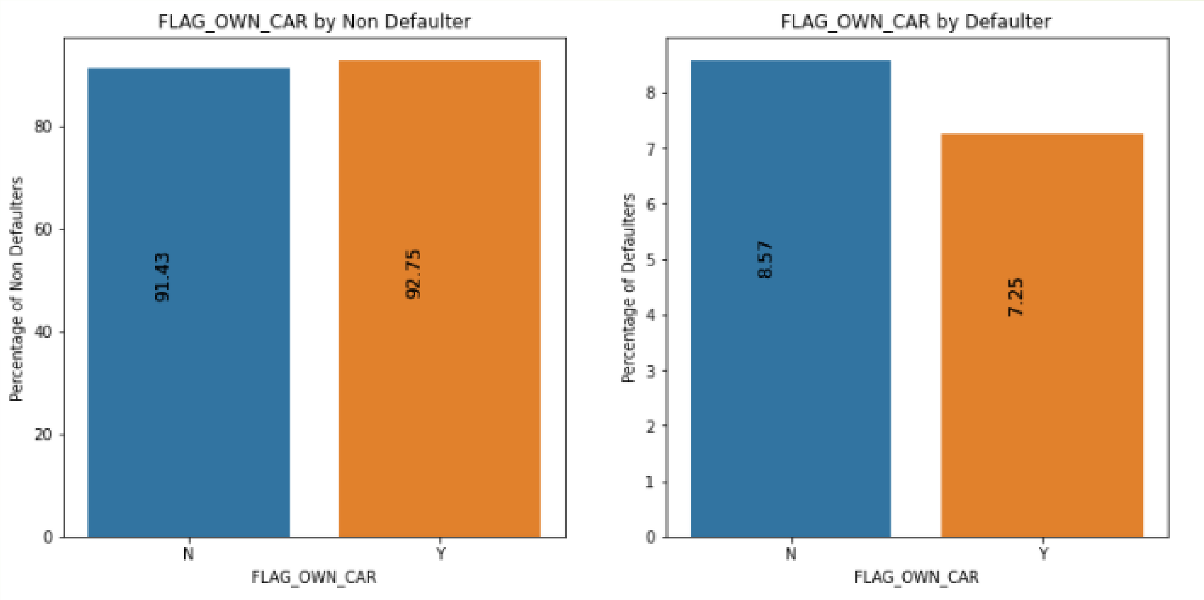
Gender	Loan Application %	Loan Application % (Non-Defaulter)	Loan Application % (Defaulter)
Male(M)	34.25%	89.81%	10.19%
Female(F)	65.75%	92.96%	7.04%

House/Car Ownership Analysis - User Demographics

Own Car % Vs Loan Applications



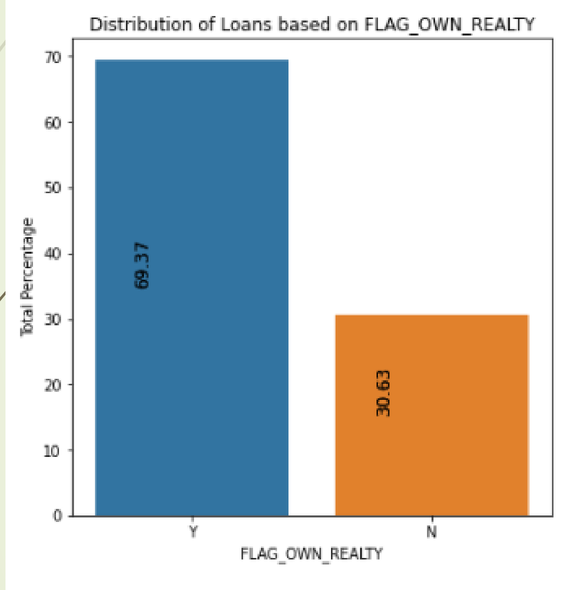
Own Car % Vs Loan Applications (Defaulter/Non-Defaulter)



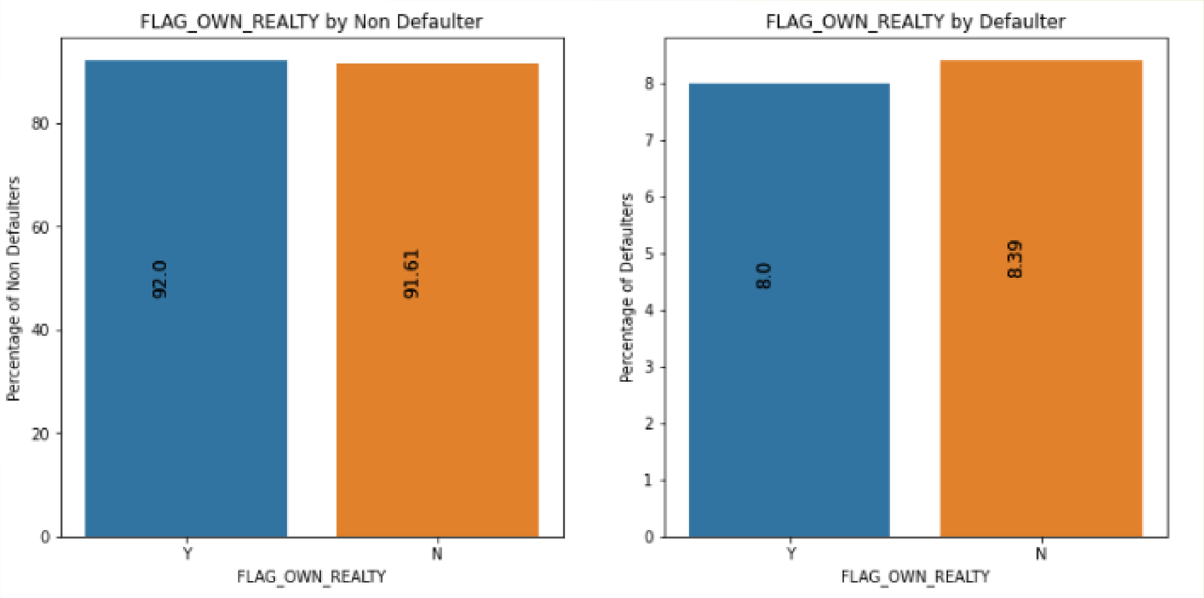
Own Car ?	Loan Application %	Loan Application % (Non-Defaulter)	Loan Application % (Defaulter)
Yes	34%	92.75%	7.25%
No	66%	91.43%	8.57%

House/Car Ownership Analysis - User Demographics (continued..)

Own House % Vs Loan Applications



Own House % Vs Loan Applications (Defaulter/Non-Defaulter)

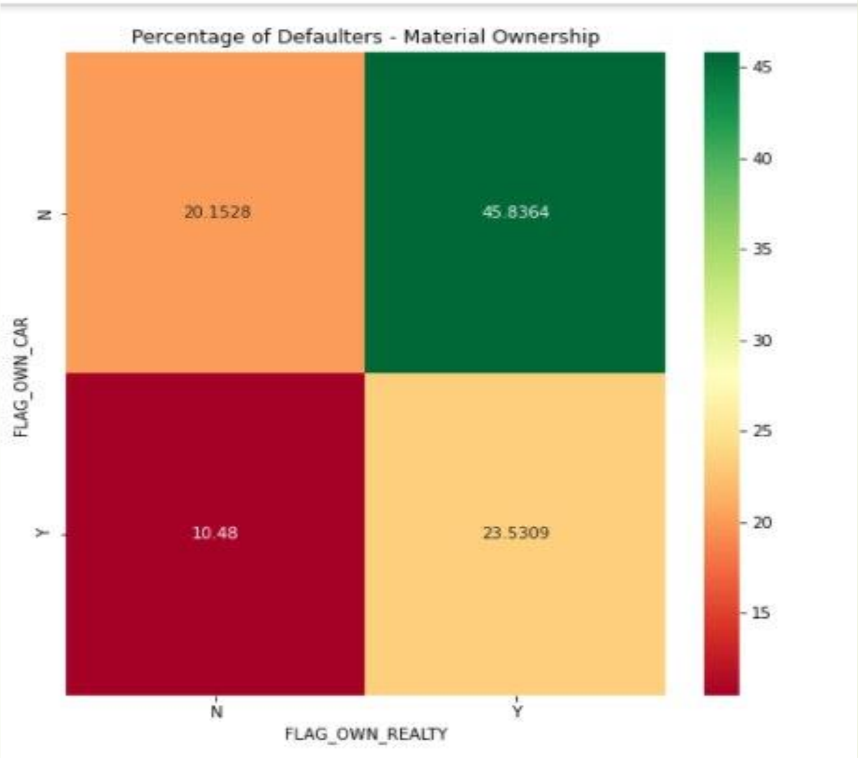


Own House ?	Loan Application %	Loan Application % (Non-Defaulter)	Loan Application % (Defaulter)
Yes	69.3%	92%	8%
No	30.7%	91.61%	8.39%

House/Car Ownership Analysis - User Demographics (continued..)

Own House/Car Vs Count of Defaulters (in %)

Own Car	Own House	Loan Application % (Non-Defaulter)
Yes	No	10.48%
No	No	20.15%
Yes	Yes	23.5%
No	Yes	45.83%

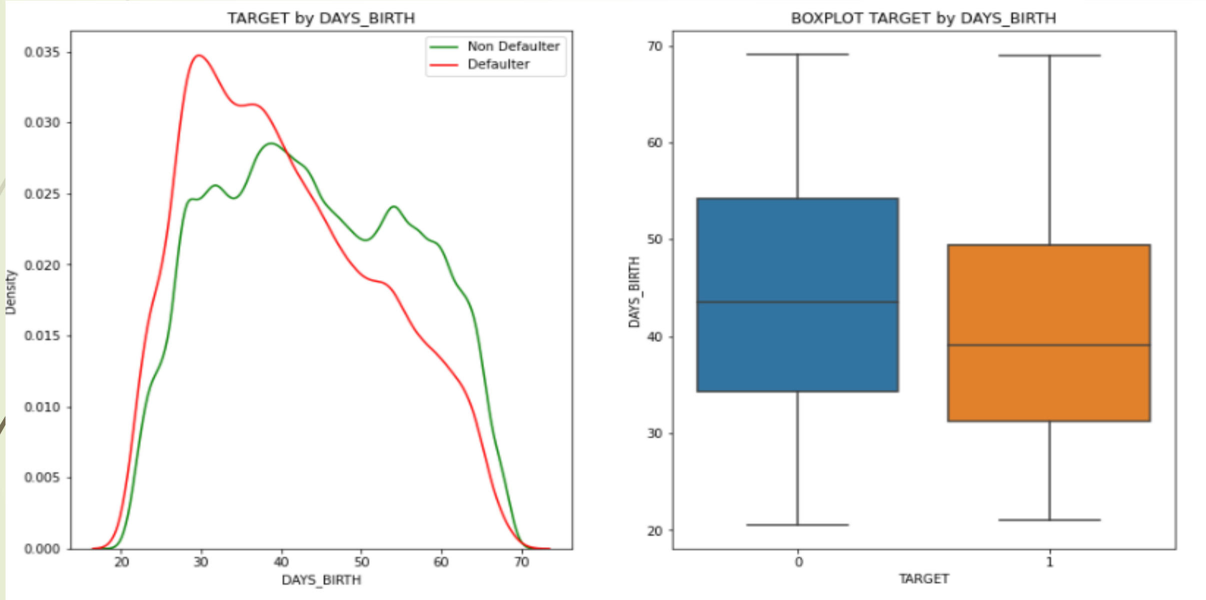


Summary:

- House and Car owned variable contains 0% null values
- Applicants who own Realty tends to apply for loan the most and default the most. It maybe because they have more commitments hence need further analysis.
- The second highest defaulters are applicants who own both House and a Car.
- Applicants who own a Car is the less likely to default.

Age Analysis - User Demographics

Age Vs Loan Applications (Defaulter/Non-Defaulter)



Target = 0 (Non-Defaulter)
Target = 1 (Defaulter)

Actions:

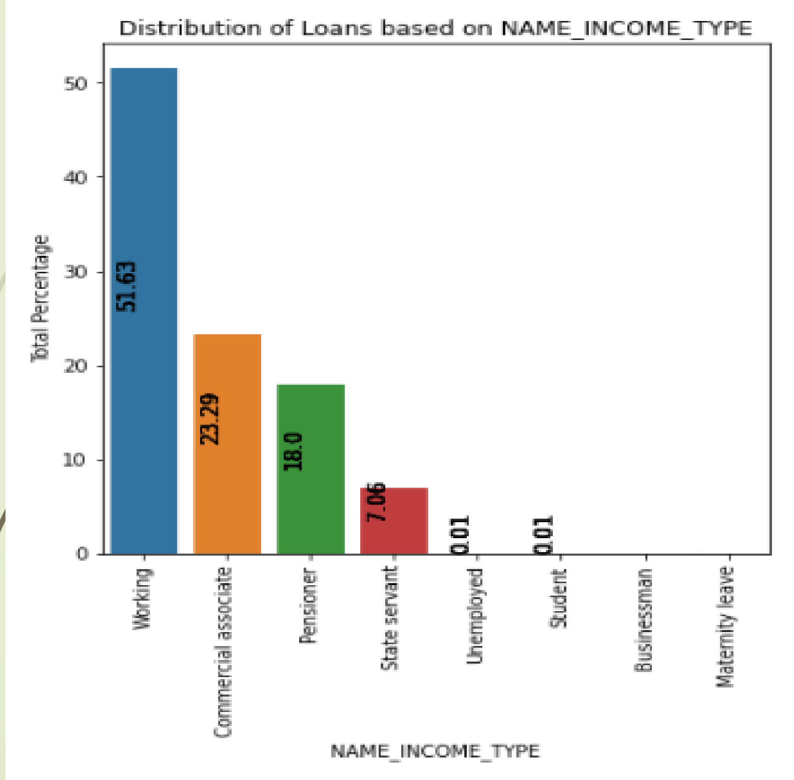
The 'DAYS_BIRTH' in the dataset was provided in days hence converting the same into Year as a new dataset column 'AGE'

Summary:

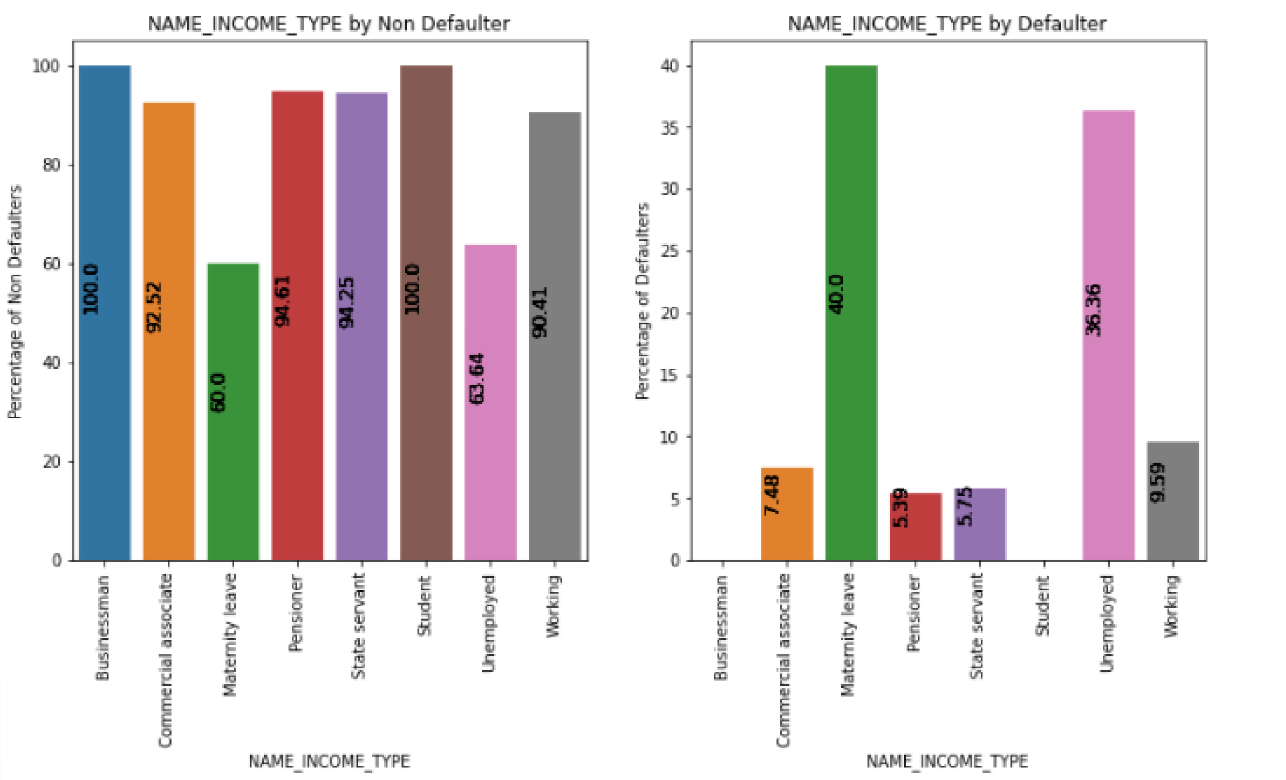
- AGE contains 0% null values
- Clients in the age group > 40 are less likely to default loans

Employment Analysis – Income Type

Income Type Vs Loan Applications



Income Type Vs Loan Applications (Defaulter/Non-Defaulter)



Employment Analysis – Income Type (continued....)

Top 3 Income Type Vs Loan Applications (%)

Income Type	Loan Application %
Working	51.63%
Commercial associate	23.29%
Pensioner	18%

Top 3 Income Type Vs Loan Applications (Defaulter)

Income Type	Loan Application %
Maternity Leave	40%
Unemployed	36.36%
Commercial Associate	7.48%

Summary:

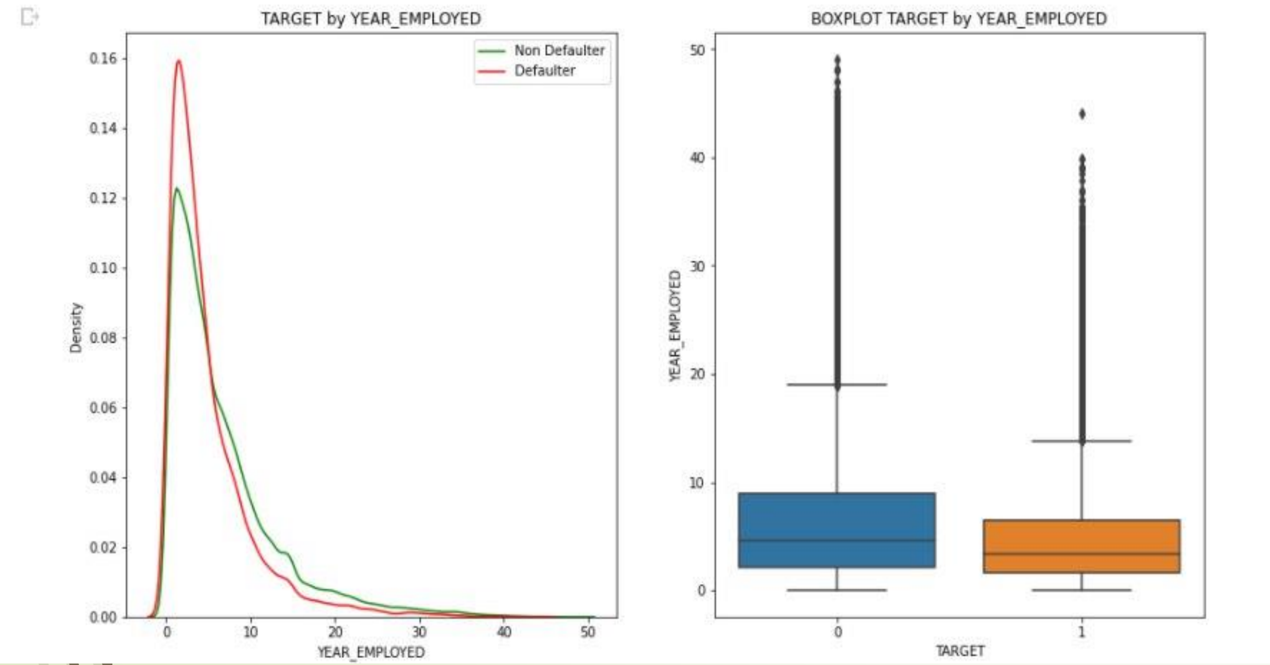
- `NAME_INCOME_TYPE` contains 0% null values
- ‘Working’ / ‘Commercial associate’ / Pensioner are the Top 3 loan applicants.
- Applicants having Income type as Maternity Leave or Unemployed are the highest defaulters though they are not in the Top 3 applicant list.
- Commercial Associates reflect in the Top 3 defaulter / Loan applicant list hence will have a larger impact if 7.48% Commercial associate defaults

Employment Analysis – Years Employed

Employed (Years) Vs Loan Applications

Analyzing Days Employed excluding 1000 years

```
[358] numerical_plot(df_app_data[df_app_data['YEAR_EMPLOYED']<1000],'YEAR_EMPLOYED')
```

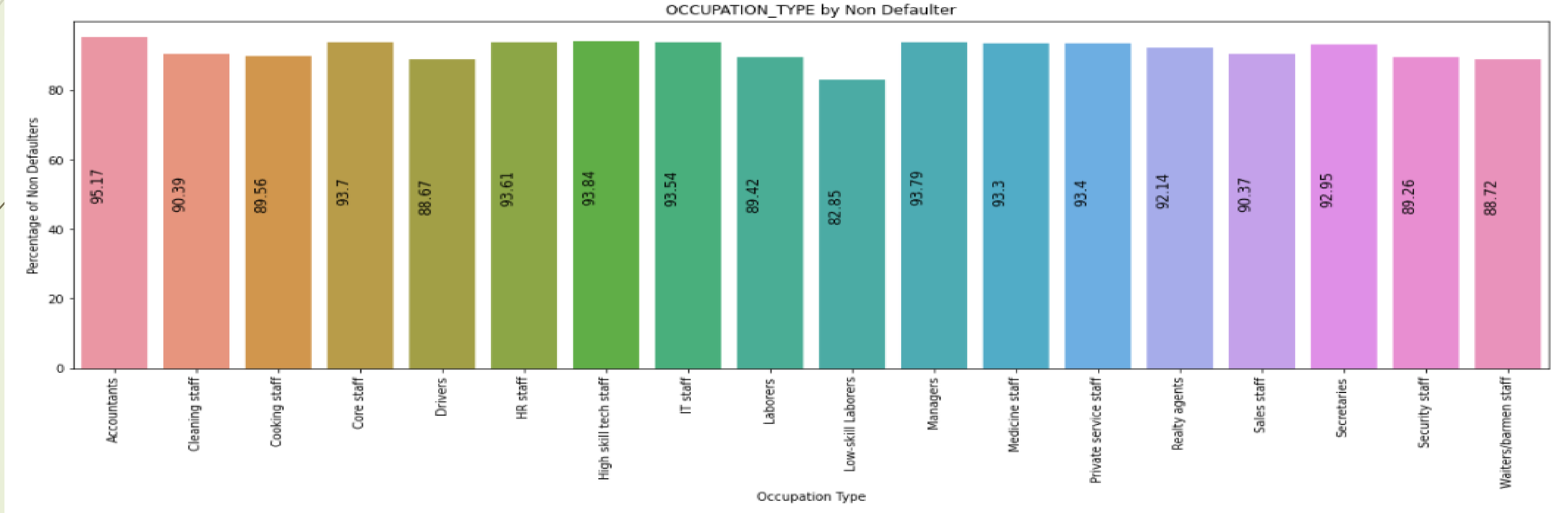


Summary:

- 'DAYS_EMPLOYED' contains 0% null values
- There are Applicants having 1000 years of experience which doesn't seem to be correct. Hence excluding the same while plotting 'DAYS_EMPLOYED'
- Applicants who have less than 8 years of experience tend to default more.

Employment Analysis – Occupation Type

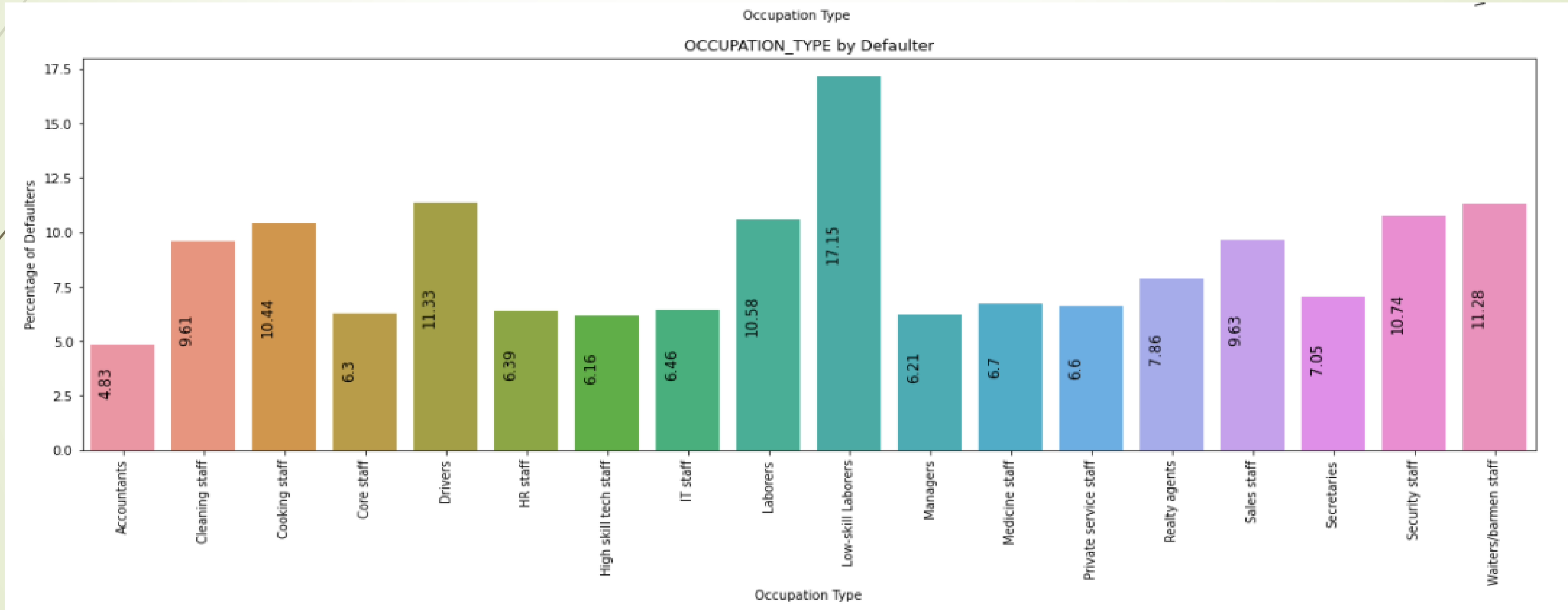
Occupation Type Vs Loan Applications (Non-Defaulter)



All the occupation Non-default % bar seems to be of the same range except "Low skill Labourers" / "Labourers" / "Cooking Staff" / "Waters/bamen staff" / "Security Staff"

Employment Analysis – Occupation Type (continued….)

Occupation Type Vs Loan Applications (Defaulter)



All the occupation Non-default % bar seems to be of the same range except "Low skill Labourers" / "Labourers" / "Cooking Staff" / "Waters/bamen staff" / "Security Staff"

Employment Analysis – Occupation Type (continued….)

Top 5
Occupation Type Vs Loan Applications (Non-Defaulter)

Occupation Type	Loan Application %
Accountants	95.17%
High Skilled Tech Staff	93.84%
Managers	93.79%
Core staff	93.70%
HR Staff	93.61%
IT Staff	93.54%

Top 5
Occupation Type Vs Loan Applications (Defaulter)

Occupation Type	Loan Application %
Low skilled Labourers	17.15%
Drivers	11.33%
Waters\bamen staff	11.28%
Security Staff	10.74%
Labourers	10.58%

Top 5
Occupation Type Vs Loan Applications

Occupation Type	Loan Application %
Laborers	31.47%
Sales staff	16.67%
Drivers	11.36%
Core staff	9.37%
Managers	7.16%

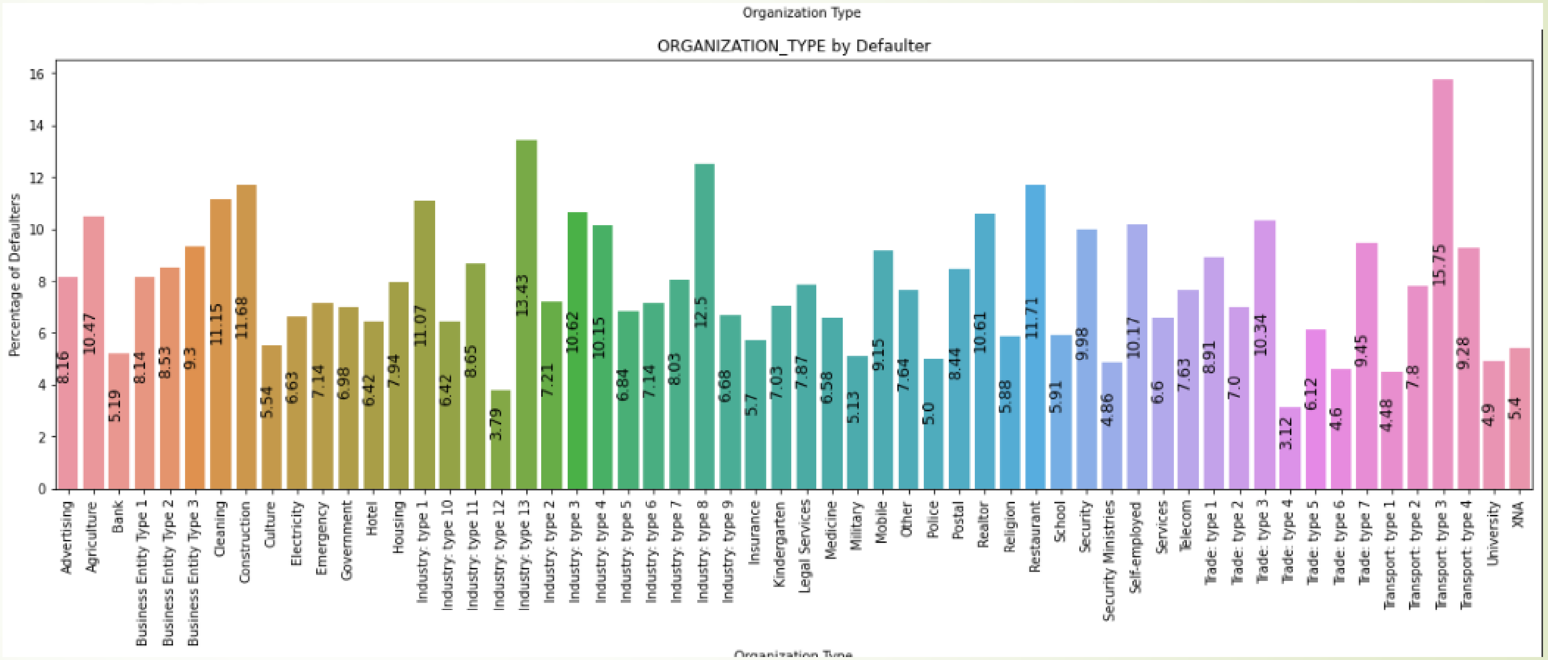
Summary:

- 'OCCUPATION_TYPE' contains 31% null values. Without a logical reason if we imputing value may introduce in an imbalance hence have not imputed the column.
- We have listed the Top5 Defaulter/ Non-Defaulter occupations for reference.
- Laborers / Drivers occupation is listed in both the Top 5 occupation types (who have applied for a loan) and have also defaulted.
- Sales staff do not appear in the Top 5 defaulter list but still responsible for 9.63% loan default

Employment Analysis – Organization Type

If we club the subcategories of Business, Industry , Trade , Transport will that depict a different picture ?

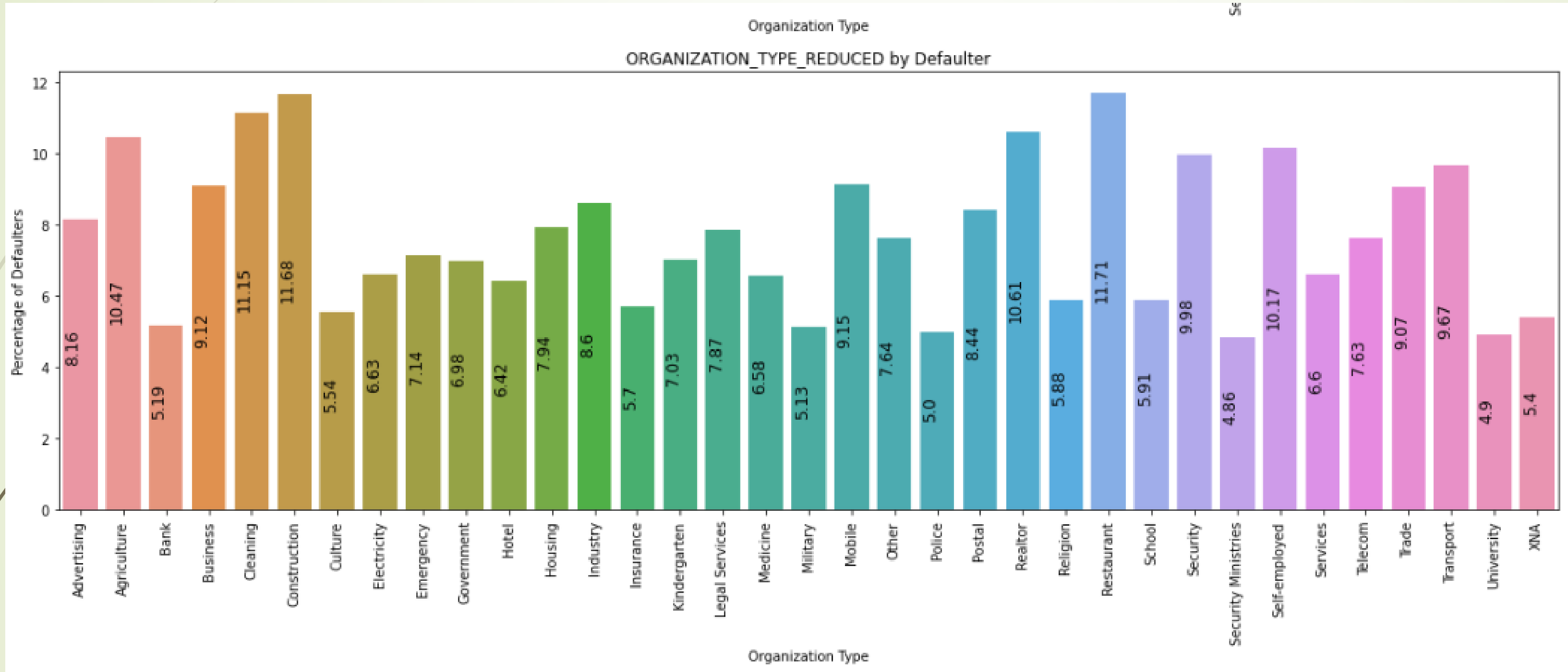
Hence will plot the same next



Summary:

- ‘ORGANIZATION_TYPE’ contains 0% null values
- Transport: type 3 ,Industry: type 13 and Industry: type 8 are more likely to default.

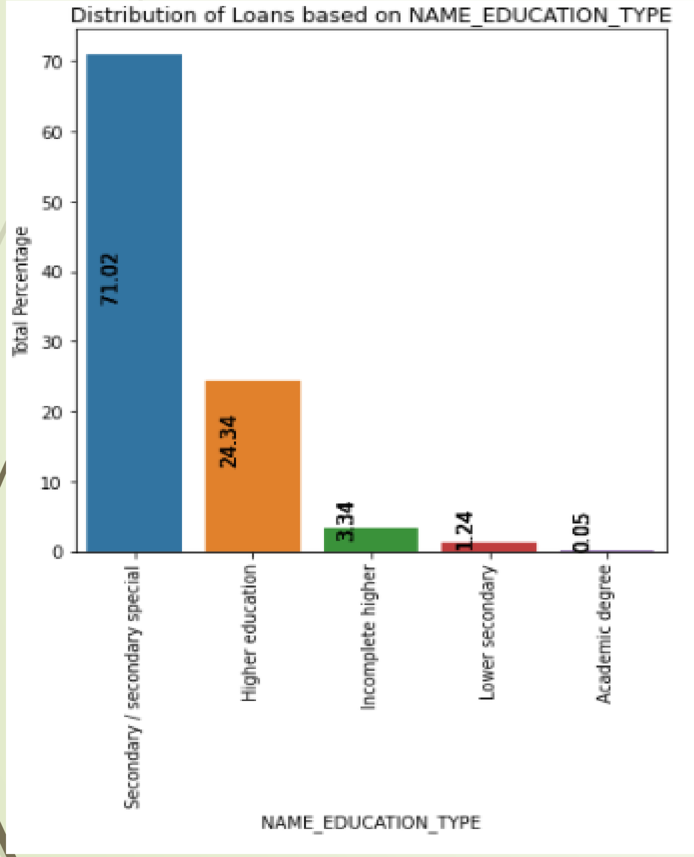
Employment Analysis – Organization Type



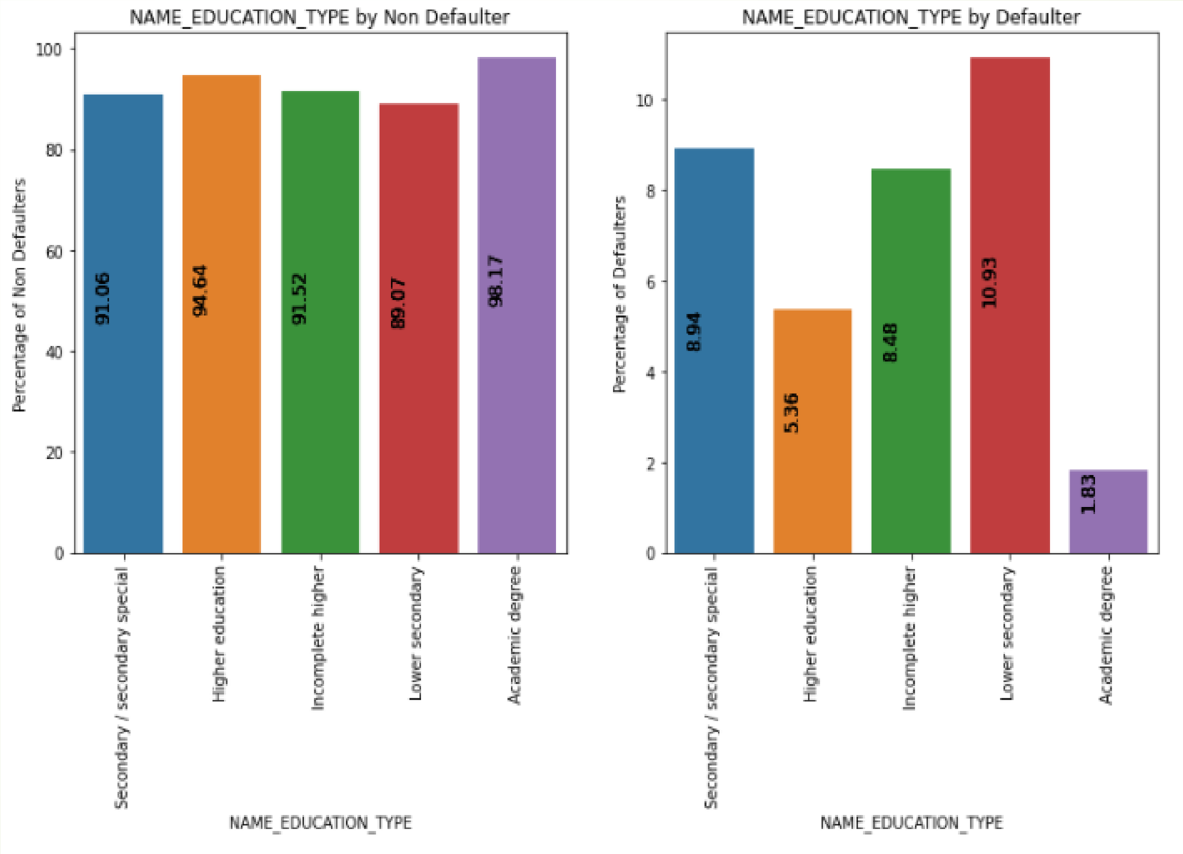
- Applicants from Restaurant, Construction and Cleaning are more likely to default which is completely different from the previous inference. Anyways it's a good inference.

Education Analysis

Education Vs Loan Application (%)



Education Vs Defaulter / Non-Defaulter



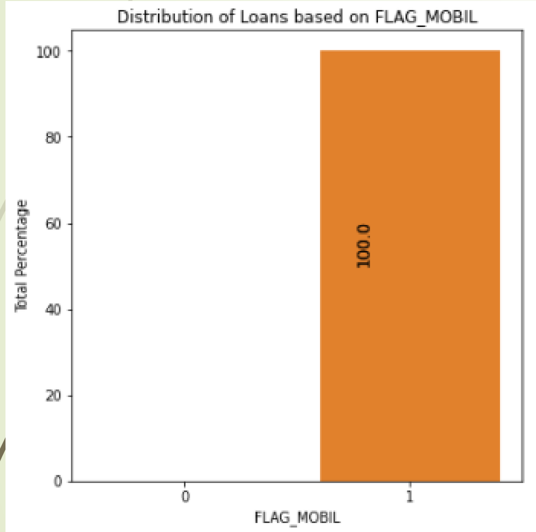
Education Analysis (continued…)

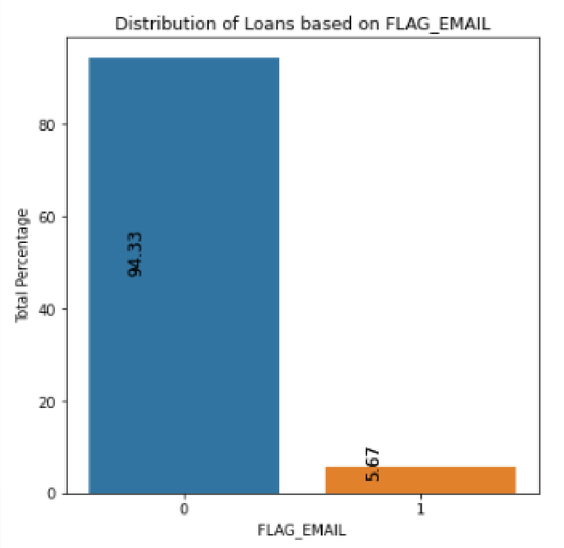
Summary:

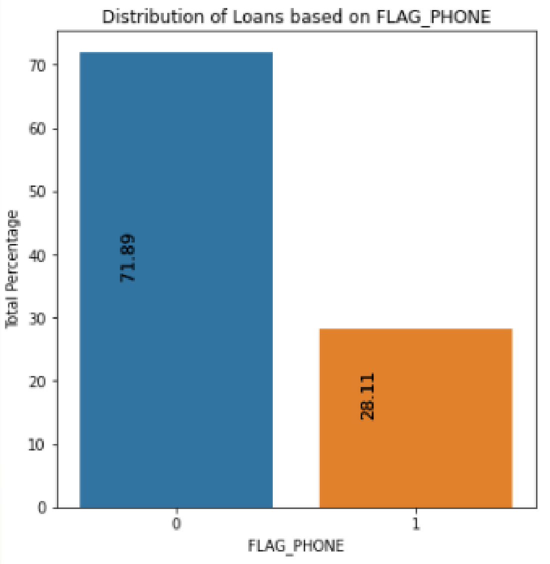
- ‘NAME_EDUCATION_TYPE’ contains 0% null values.
- Applicants who are having highest education as Secondary or less seems to have applied or defaulted the most.
- There is a ‘Incomplete’ category as well which default, but the column of such Applicant is less.
- It seems with more education, the applicants are less likely t default

Contact Rating Analysis

Percentage (%) of loan applicants who provided individual Email / Mobile / Phone details in the application



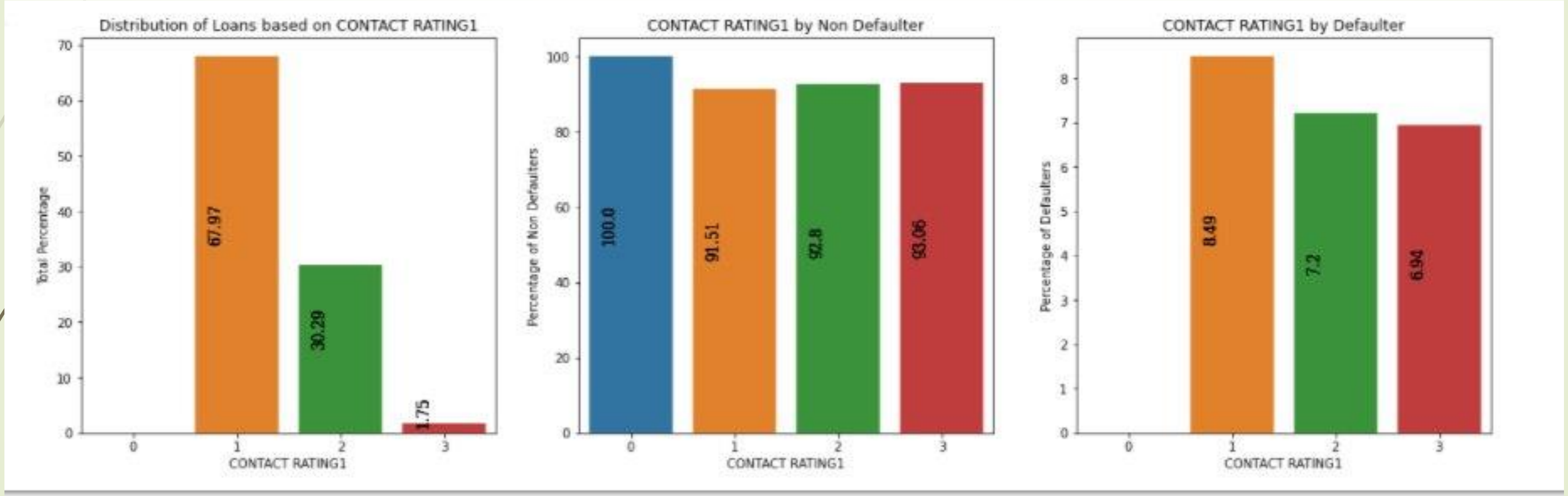




- 100% Applicants provide mobile details in their loan application.
- 94.33% applicants do not provide their email address.
- 71.89% do not provide their home phone number

Contact Rating Analysis (continued…)

Percentage (%) of loan applicants who provided Email / Mobile / Phone details in the application

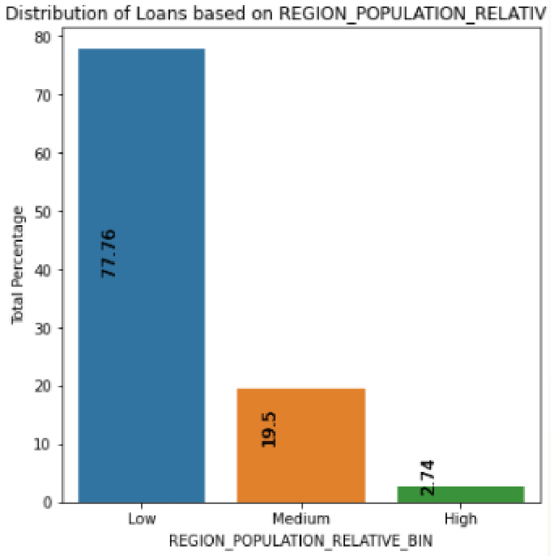


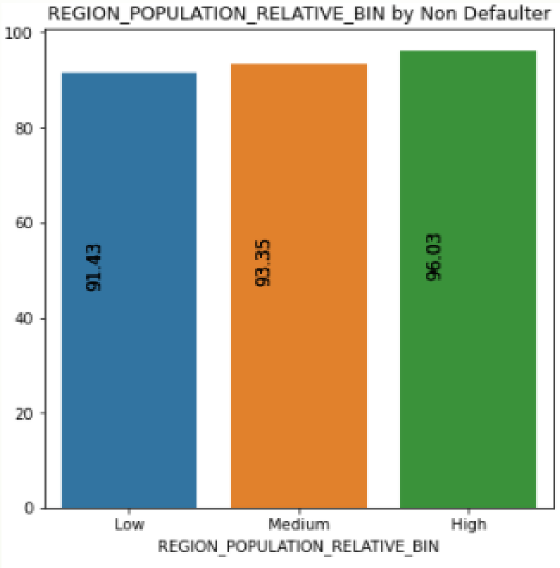
- Majority applicants provide one contact details during their application.
- With at least 2 or more contact details applicants will be less likely (<= than 7.2%) to default.

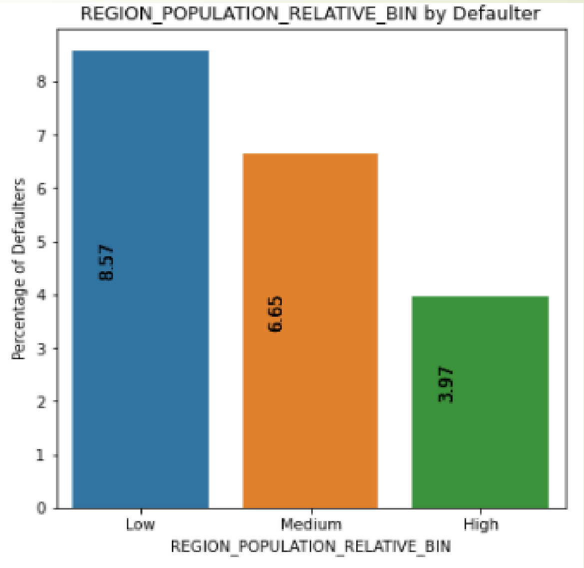
Dwelling Rating

Actions:

❖ ‘REGION_POPULATION_RELATIVE’ is binned into three categories ‘Low’, ‘Medium’ and ‘High’ for better visibility.



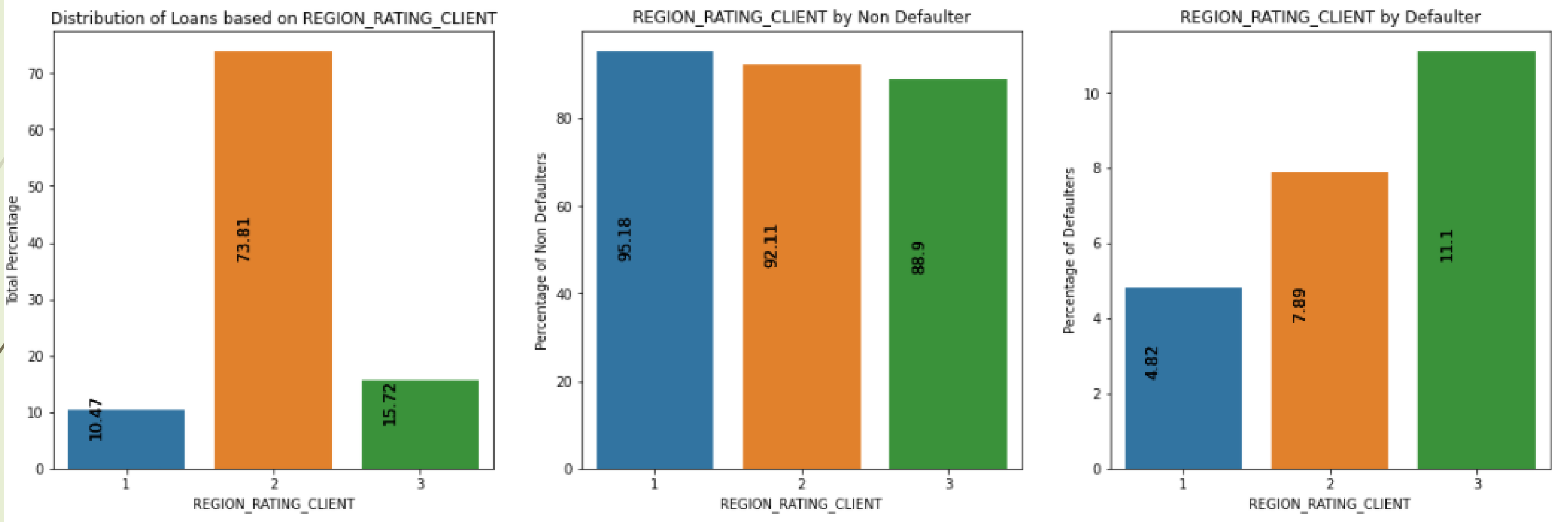




Region Population Category	Overall (%)	Non-Defaulters(%)	Defaulters(%)
Low	77.76%	91.43%	8.57%
Medium	19.5%	93.35%	6.65%
High	2.74%	96.03%	3.97%

Dwelling Rating (continued)

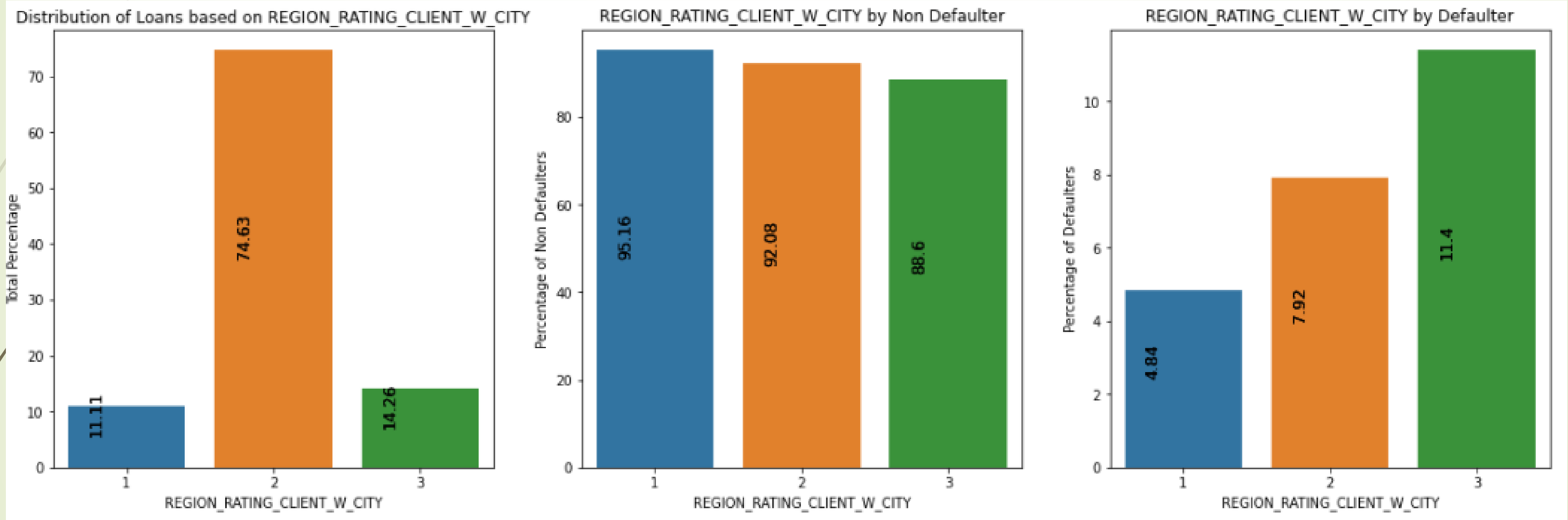
- 'REGION_RATING_CLIENT' analysis.



Region Client Rating	Overall (%)	Non-Defaulters(%)	Defaulters(%)
1	10.47%	95.18%	4.82%
2	73.81%	92.11%	7.89%
3	15.72%	88.9%	11.1%

Dwelling Rating (continued)

- REGION_RATING_CLIENT_W_CITY' analysis.



Region Client Rating (City)	Overall (%)	Non-Defaulters(%)	Defaulters(%)
1	11.11%	95.16%	4.84%
2	74.63%	92.08%	7.92%
3	14.26%	88.6%	11.4%

Dwelling Rating (continued)

Summary:

- 'REGION_POPULATION_RELATIVE', 'REGION_RATING_CLIENT_W_CITY','NAME_HOUSING_TYPE' & 'REGION_RATING_CLIENT contains 0% null values.
- Majority Applicants are from relatively low populated area and are more likely to default.
- From Client Rating perspective majority applicants are from people having rating 2 and is the second most defaulter group.
- Though the applicants having 3 client rating tops the defaulter list the overall application volume is less compared to rating 2.

Region Population Category	Overall (%)	Non-Defaulters(%)	Defaulters(%)
Low	77.76%	91.43%	8.57%
Medium	19.5%	93.35%	6.65%
High	2.74%	96.03%	3.97%

Region Client Rating	Overall (%)	Non-Defaulters(%)	Defaulters(%)
1	10.47%	95.18%	4.82%
2	73.81%	92.11%	7.89%
3	15.72%	88.9%	11.1%

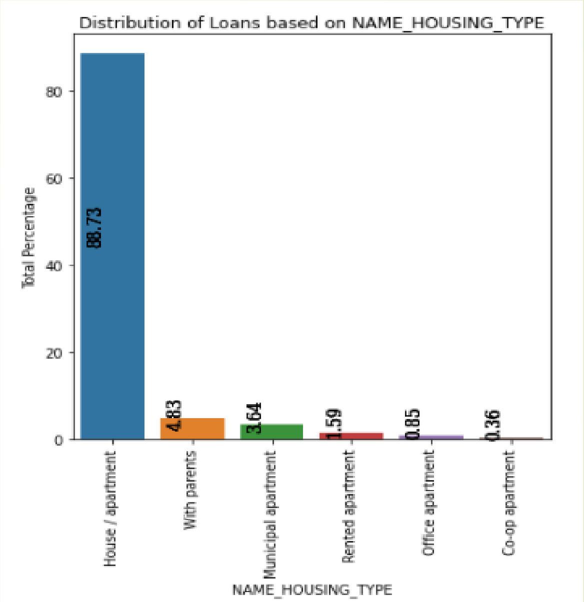
Region Client Rating (City)	Overall (%)	Non-Defaulters(%)	Defaulters(%)
1	11.11%	95.16%	4.84%
2	74.63%	92.08%	7.92%
3	14.26%	88.6%	11.4%

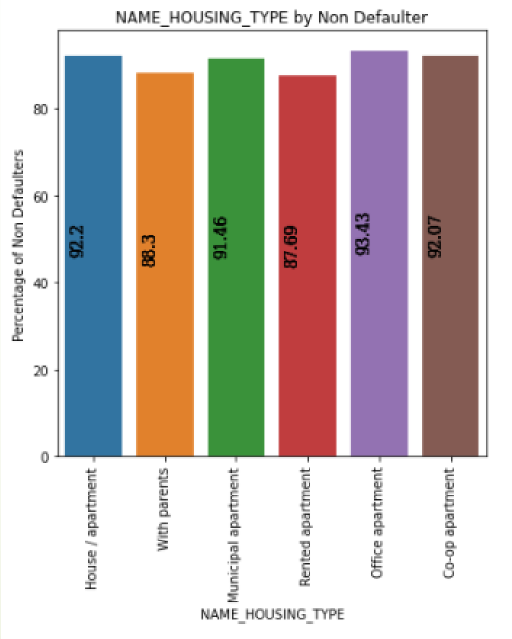
Dwelling Rating (continued)

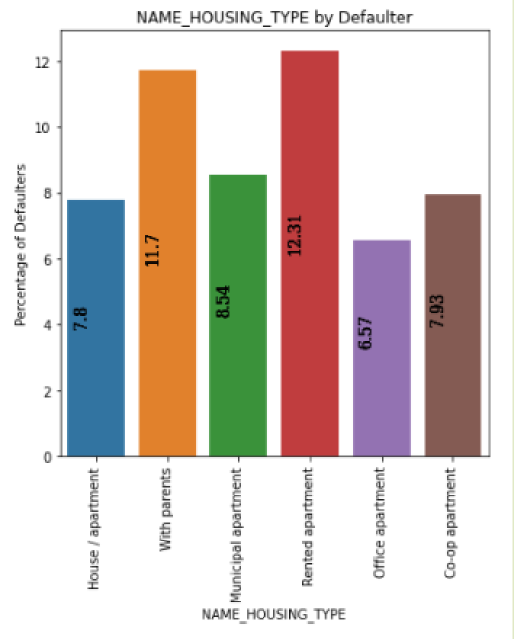
Housing Type	Overall (%)	Non-Defaulters(%)	Defaulters(%)
House / apartment	88.73%	92.2%	7.8%
Wit Parents	4.83%	88.3%	11.7%
Municipal Apartment	3.64%	91.46%	8.54%
Rented Apartment	1.59%	87.69%	12.31%
Office Apartment	0.85%	93.43%	6.57%
Co-op apartment	0.36%	92.07%	7.93%

Summary:

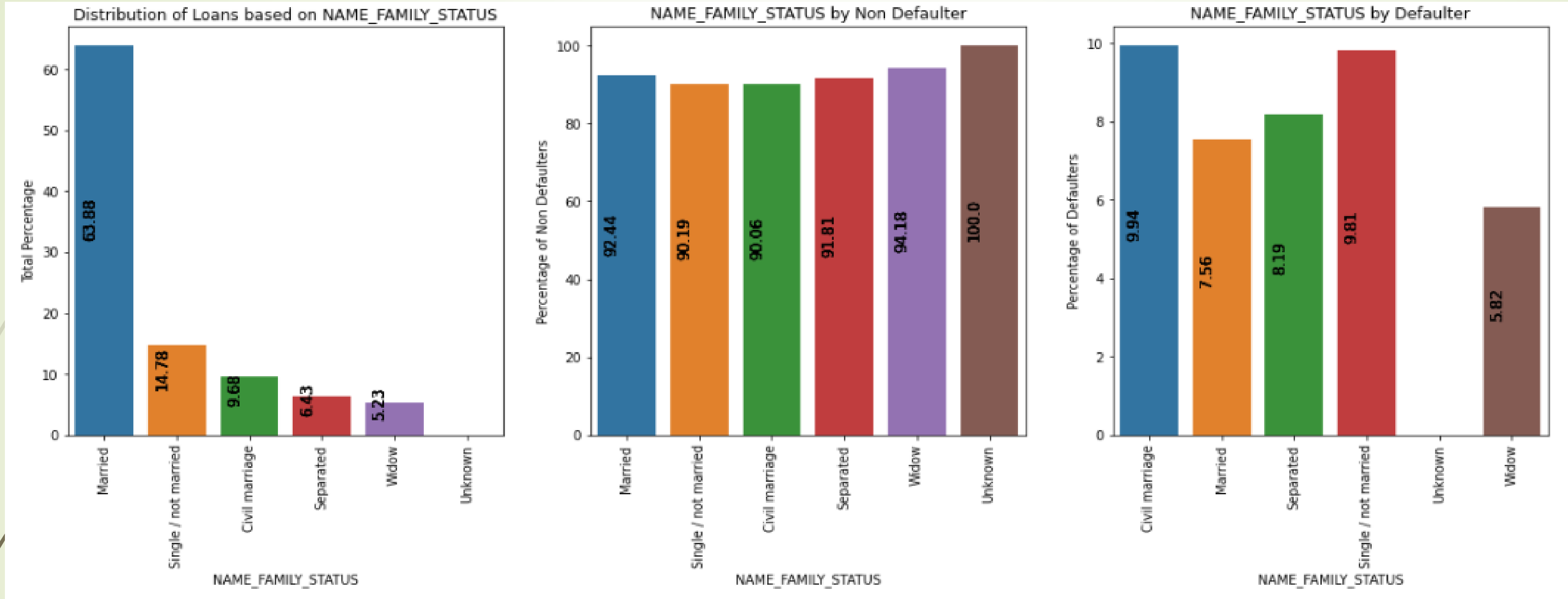
- Applicants staying in a Rented apartment / With Parents / Municipal apartment is more likely to default
- Applicants staying in House/Apartment are the major loan applicants with 7.8% defaulters.







Marital Status Analysis



Marital Status	Overall (%)	Non-Defaulters(%)	Defaulters(%)
Married	63.88%	92.44%	7.56%
Single/Not Married	14.78%	90.19%	9.81%
Civil Marriage	9.68%	90.06%	9.94%
Separated	6.43%	91.81%	8.19%
Widow	5.23%	94.18%	5.82%

Marital Status Analysis (continued)

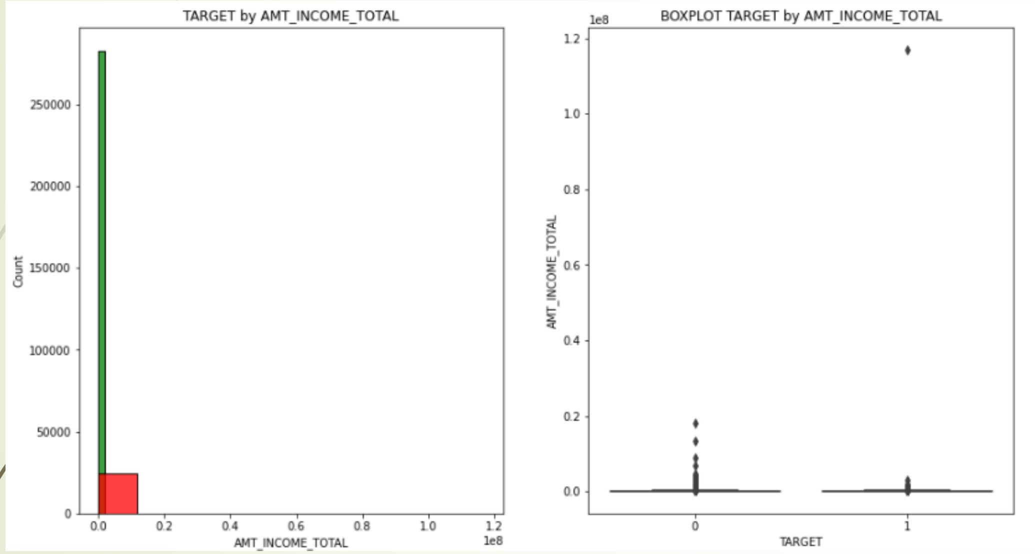
Summary:

- ‘NAME_FAMILY_STATUS’ contains 0% null values.
- Majority of the applicants (63.88%)who have married have applied for a loan.
- Civil Marriage & Single/not married family status applicants top the defaulter list followed by Separated and Married category.
- Since Marriage applicant category column is significant the defaulters will be quite significant.

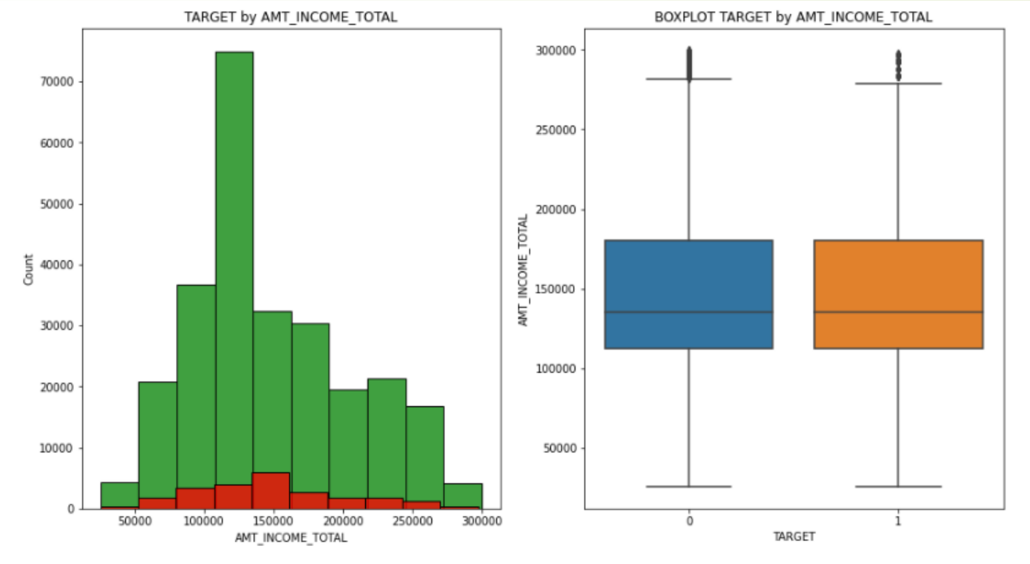
Marital Status	Overall (%)	Non-Defaulters(%)	Defaulters(%)
Married	63.88%	92.44%	7.56%
Single/Not Married	14.78%	90.19%	9.81%
Civil Marriage	9.68%	90.06%	9.94%
Separated	6.43%	91.81%	8.19%
Widow	5.23%	94.18%	5.82%

Income Analysis

Applicant Income distribution (with outliers)



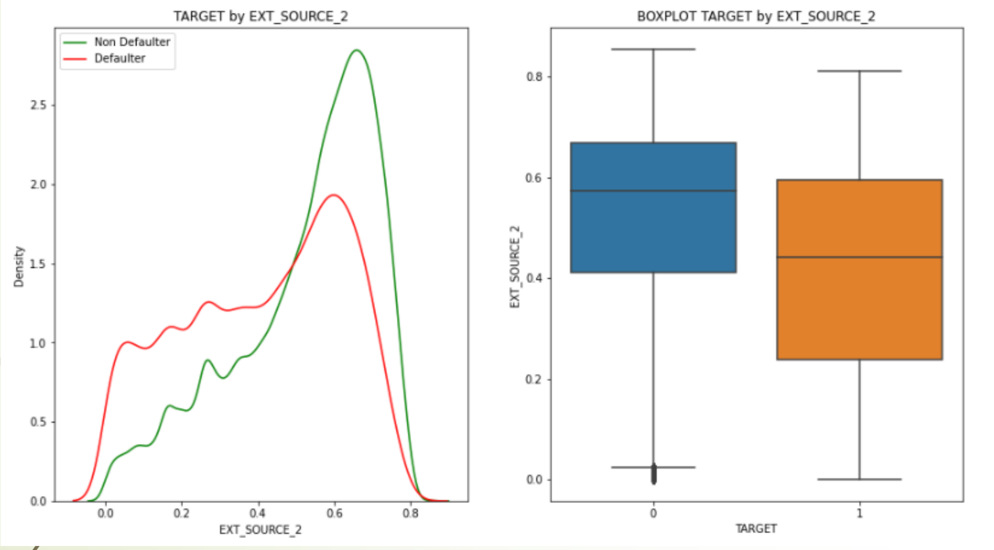
Applicant Income distribution (without outliers)

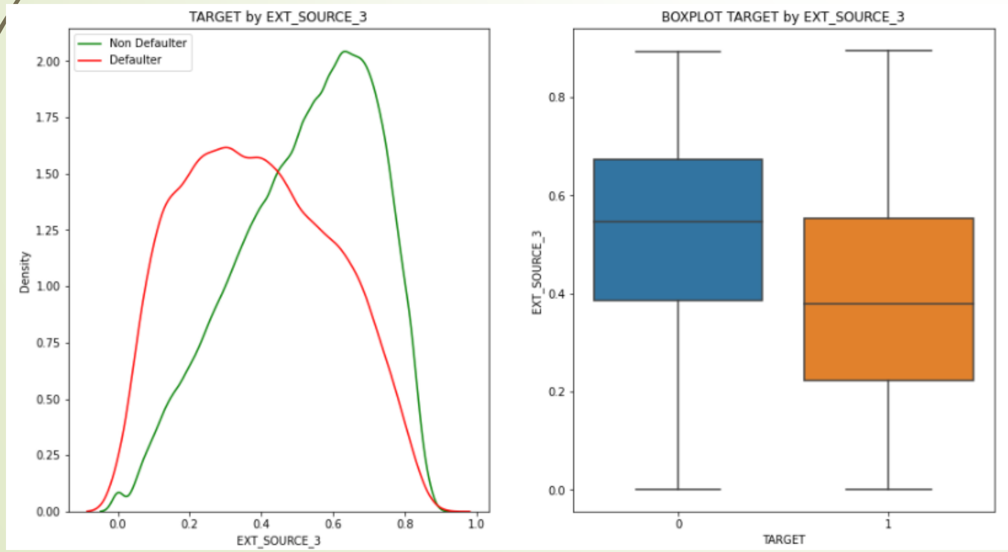


Summary:

- 'AMT_INCOME_TOTAL' contains 0% null values.
- The variable contained some outliers hence we removed the outliers
- Majority applicant are of the Income Group 100k-150k.
- Every category have some % of defaulters/non-defaulters there its difficult to conclude if any income category applicants are good. Though income range between 110k – 140k looks promising

User Rating Analysis

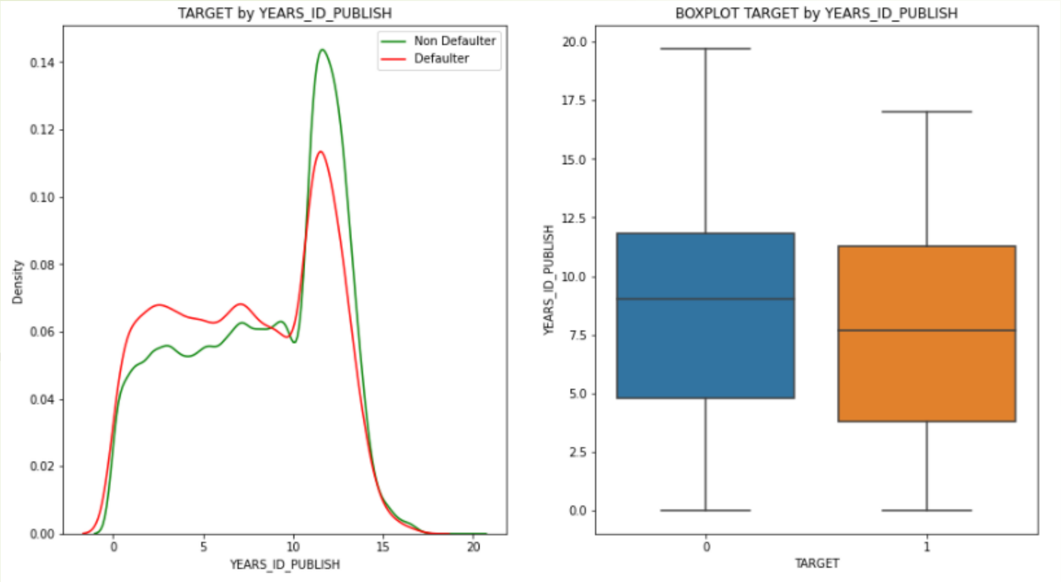


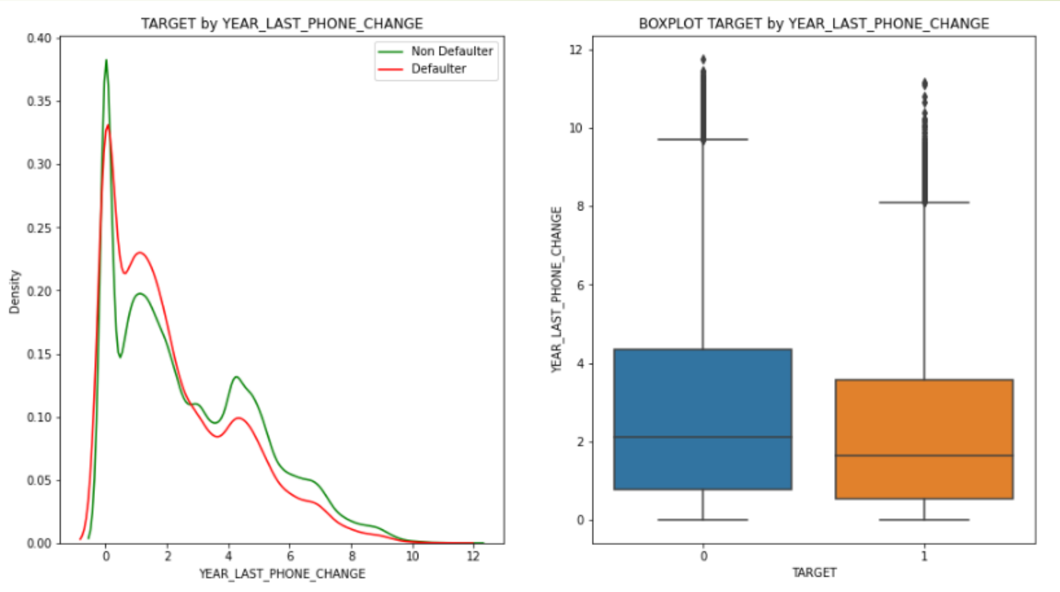


Summary:

- 'EXT_SOURCE_2' variable contains 0.21% null values whereas 'EXT_SOURCE_3' contains 19.8% null values. Since the null content of 'EXT_SOURCE_2' is not much hence imputing the value with median value
- From both the variables it seems that if the rating for the user is less than around 0.55% they are more likely to default
- The boxplot clearly shows that Non-Defaulters have a better user rating compared to Defaulters

ID/Phone Change Analysis

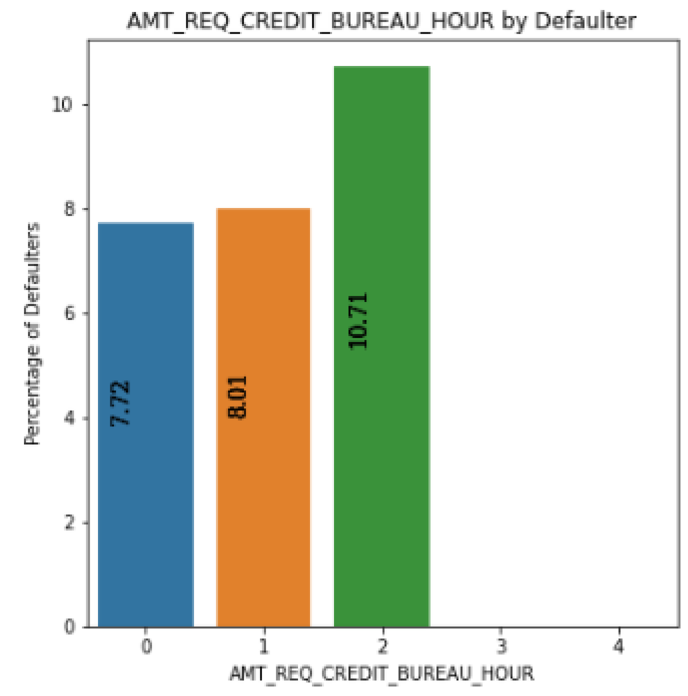


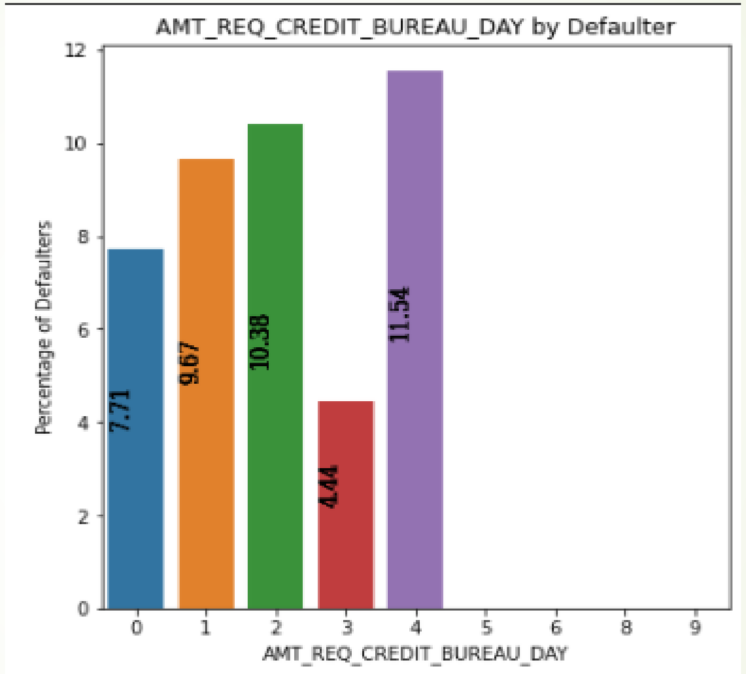


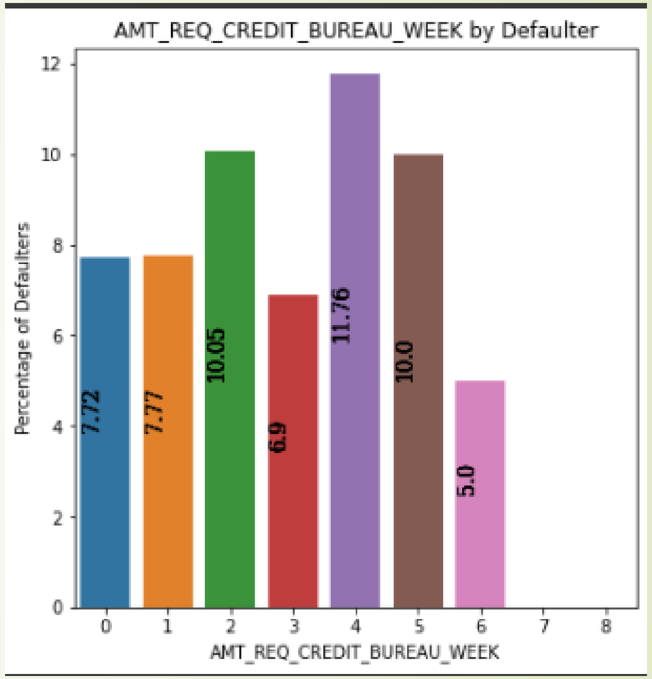
Summary:

- 'DAYS_ID_PUBLISH' variable contains 0% null values whereas 'DAYS_LAST_PHONE_CHANGE' contains 0.03% null values. We will look into the data without imputing any rating values.
- Changed the days into Years
- Applicants who have changed their ID < 7-8 years and have changed their phone numbers < 2.5 years is more likely to default.

Credit Bureau Check Analysis



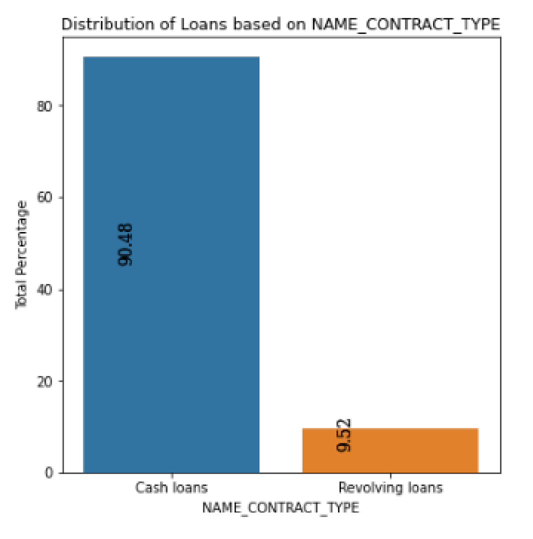


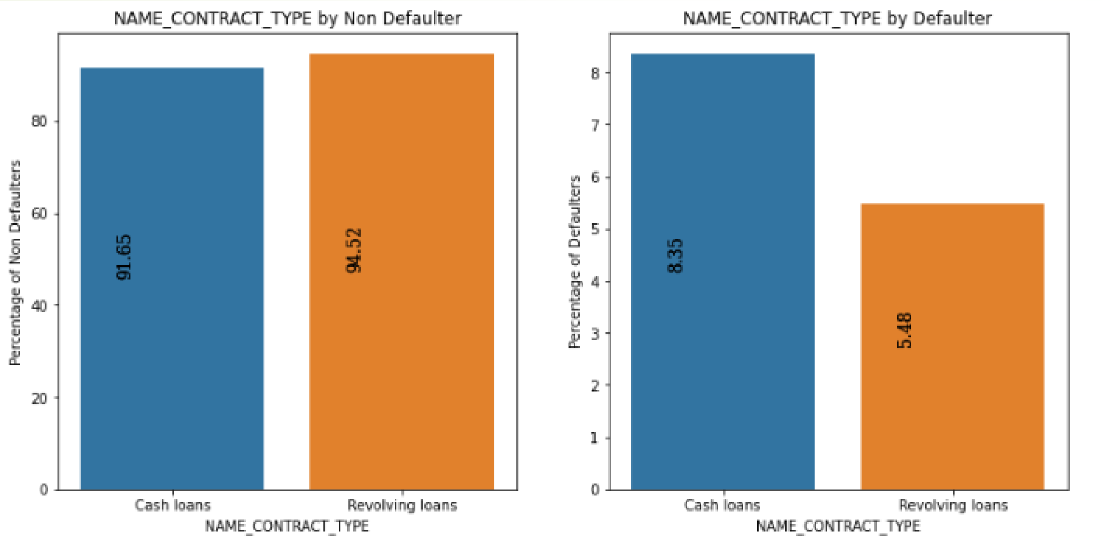


Summary:

- Its difficult to quantify the extract number of Credit Bureau requests to tag any applicant to be default but from the data it's seems if an applicant enquires more about their rating in a day, they are more likely to default with some exceptions like 3/5/6 enquires in a day/week.

Loan Summary – Loan Type

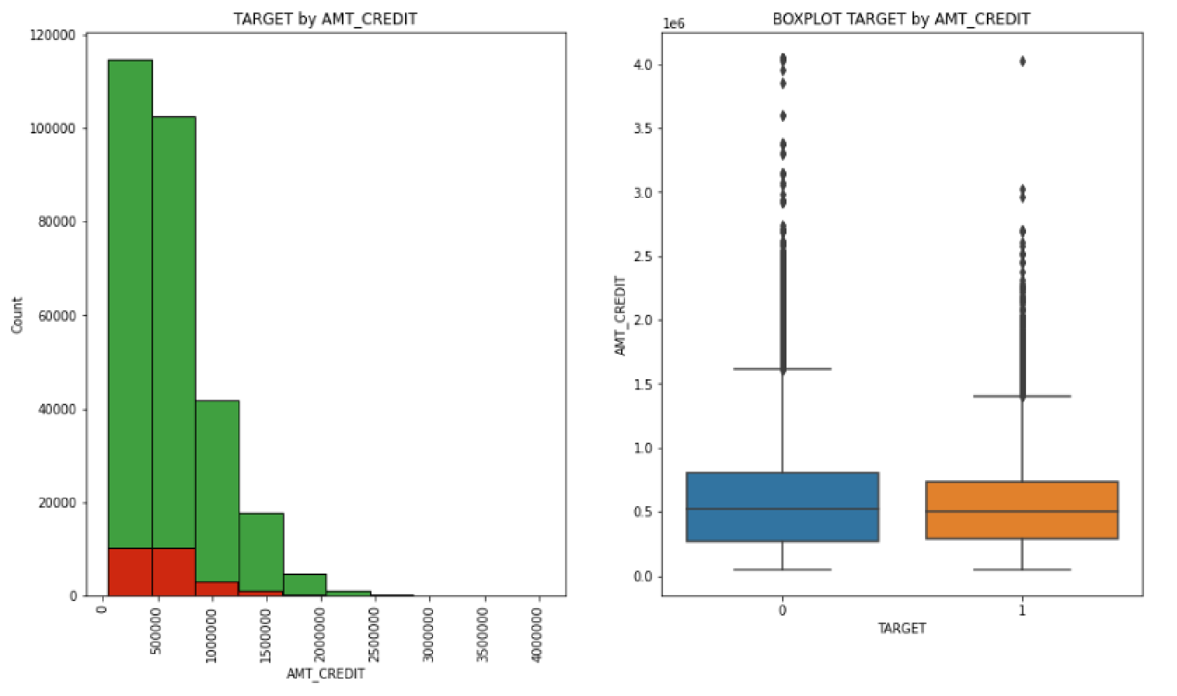




Summary:

- 90.48% & 9.52% applicants applied for cash loan and revolving loans, respectively.
- Applicants who availed cash loan seems to default more compared to the revolving loans.

Loan Summary – Amount Credited

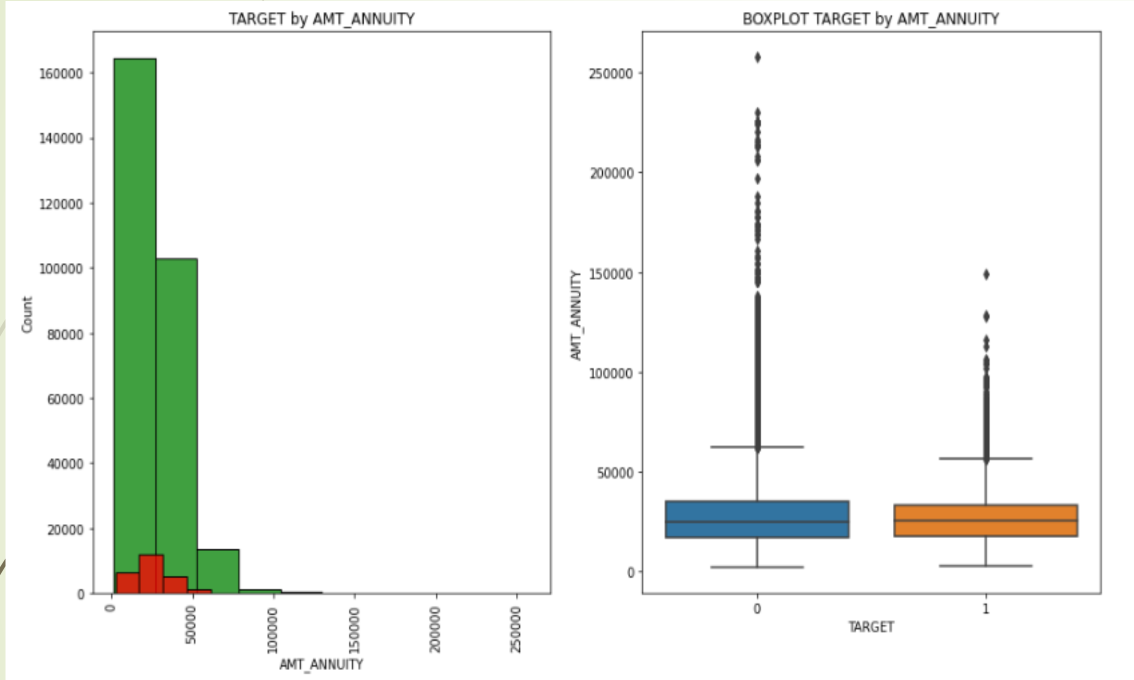


AMT_CREDIT_BIN	DEFAULT	NON DEFAULT	TOTAL	% defaulters	% non defaulters
10K-50K	23	538	561	4.0	96.0
50K-100K	308	5135	5443	6.0	94.0
100K-150K	762	11393	12155	6.0	94.0
150K-200K	1397	16588	17985	8.0	92.0
200K-300K	4321	50492	54813	8.0	92.0
300K-400K	2623	23715	26338	10.0	90.0
400K-500K	3171	28867	32038	10.0	90.0
500K-600K	3523	30709	34232	10.0	90.0
600K-700K	2097	21952	24049	9.0	91.0
700K-800K	1466	17727	19193	8.0	92.0
800K-900K	1555	20237	21792	7.0	93.0
900K-1M	647	8280	8927	7.0	93.0
1M+	2932	47053	49985	6.0	94.0

Summary:

- AMT_CREDIT contains 0% null values.
- Majority amount credited < 500K followed by 500K – 1M+
- Applicants got who the credit amount was < 50K is less likely to default followed by other categories as depicted in the table above

Loan Summary –Amount Annuity

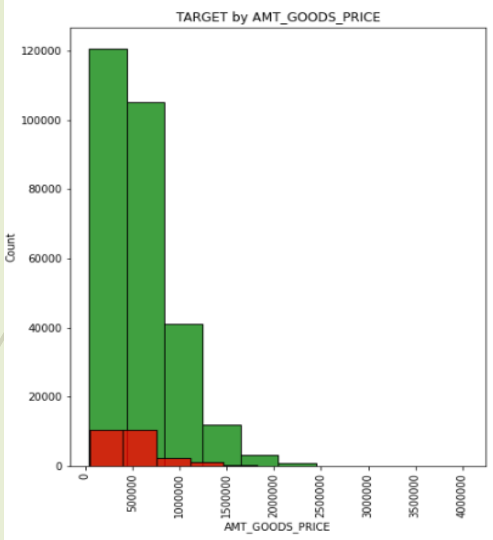


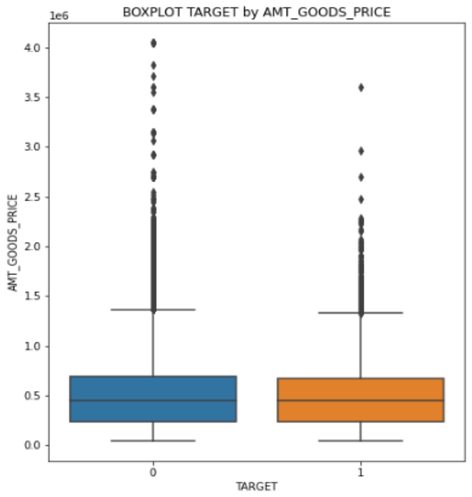
AMT_ANNUITY_BIN	DEFAULT	NON DEFAULT	TOTAL	% defaulters	% non defaulters
10K-20K	6434	76788	83222	8.0	92.0
20K-30K	8258	84220	92478	9.0	91.0
30K-40K	5251	53311	58562	9.0	91.0
40K-50K	2009	26660	28669	7.0	93.0
50K-80K	1178	18596	19774	6.0	94.0
380K-100K	35	983	1018	3.0	97.0
100K+	10	459	469	2.0	98.0

Summary:

- AMT_ANNUITY contains few null values, hence imputing the values with the median annuity amount.
- Majority amount Annuity is less than < 50K
- No significant Defaulter % observed from the Annuity table. With Annuity amount the % of defaulter decreases. It may be because of the number of loans of that Annuity value is less.

Loan Summary – Amount Goods Price



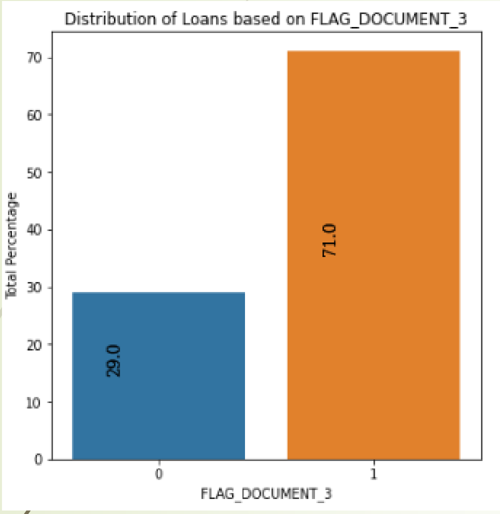


AMT_GOODS_PRICE_BIN	DEFAULT	NON DEFAULT	TOTAL	% defaulters	% non defaulters
10K-50K	59	1268	1327	4.0	96.0
50K-100K	467	6915	7382	6.0	94.0
100K-150K	1100	14844	15944	7.0	93.0
150K-200K	1434	15578	17012	8.0	92.0
200K-300K	5283	57478	62761	8.0	92.0
300K-400K	2230	18989	21219	11.0	89.0
400K-500K	5923	51328	57251	10.0	90.0
500k+	8329	116286	124615	7.0	93.0

Summary:

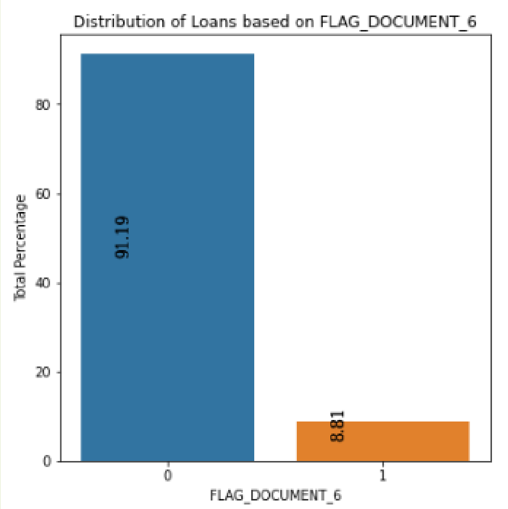
- Goods price variable had 0.09% null values. The null value is for Revolving loans only. Therefore, based on further analysis we imputed the values with AMT_CREDIT instead of imputing with the column median value. This is based on the fact, that the bank will only credit amount based on the purpose/purchase value shown in the loan.
- The histogram plot seems to be displaying the same inference as of the AMT_CREDIT. Majority amount credited < 500K followed by 500K to 1M+
- Applicants for whom the Good's Price is < 50< is having the least defaulters.

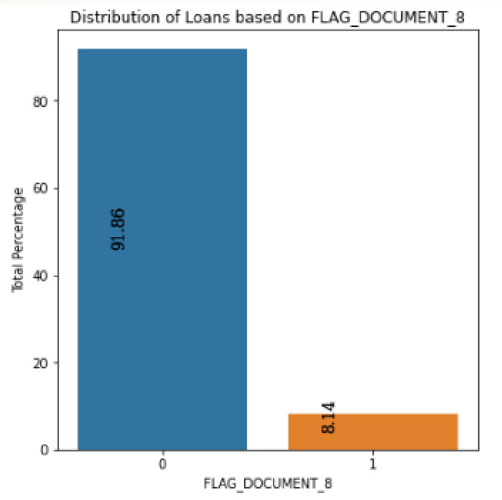
Loan Summary – Documents

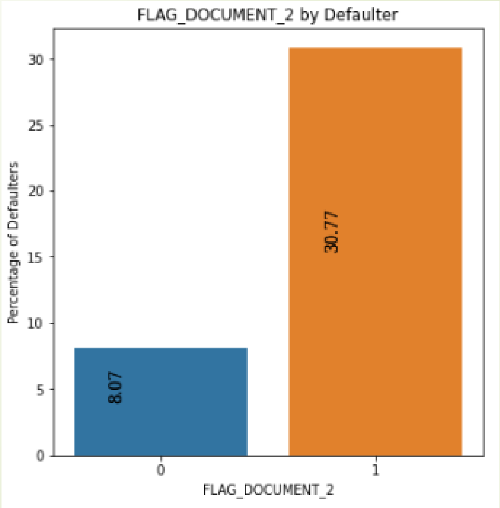


Summary:

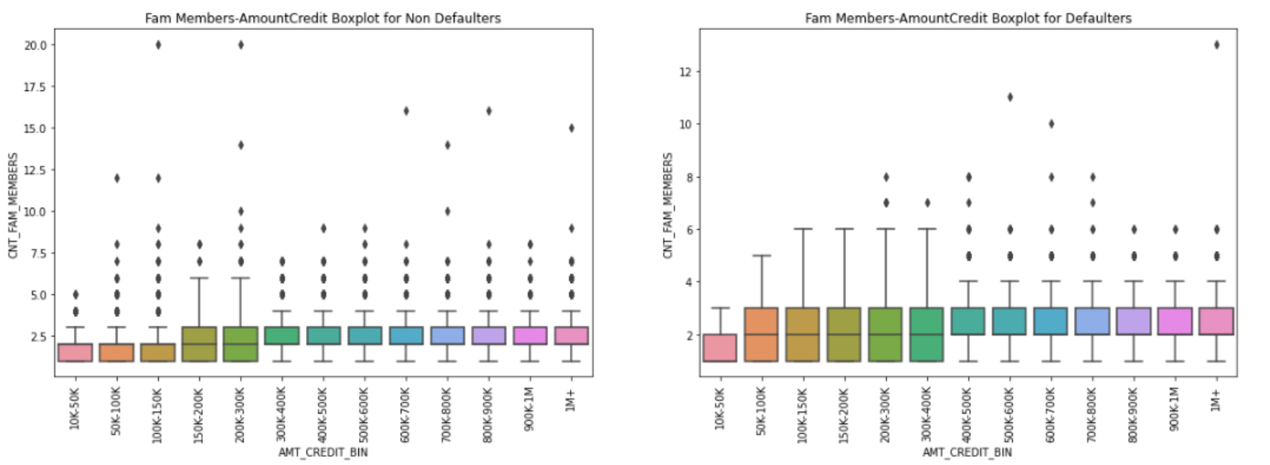
- Out of all the 21 documents only 71%, 8.81% and 8.14% applicants have provided Document 3, 6 and 8 respectively. Rest all documents were not provided by the applicants.
- Applicants who have provided document_2 seems to have defaulted the most

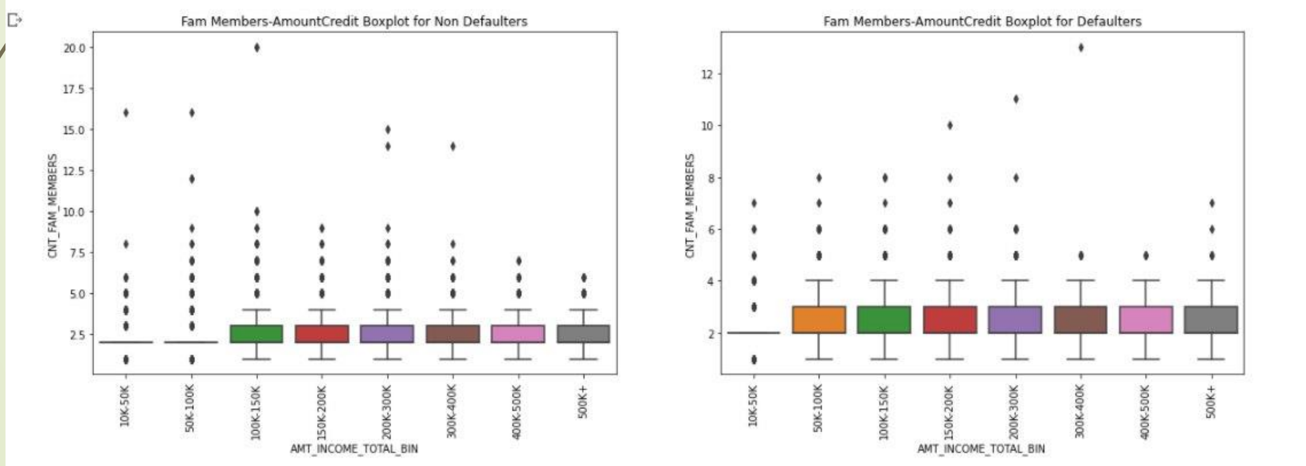






Family, Income, Amount Credit Vs Target

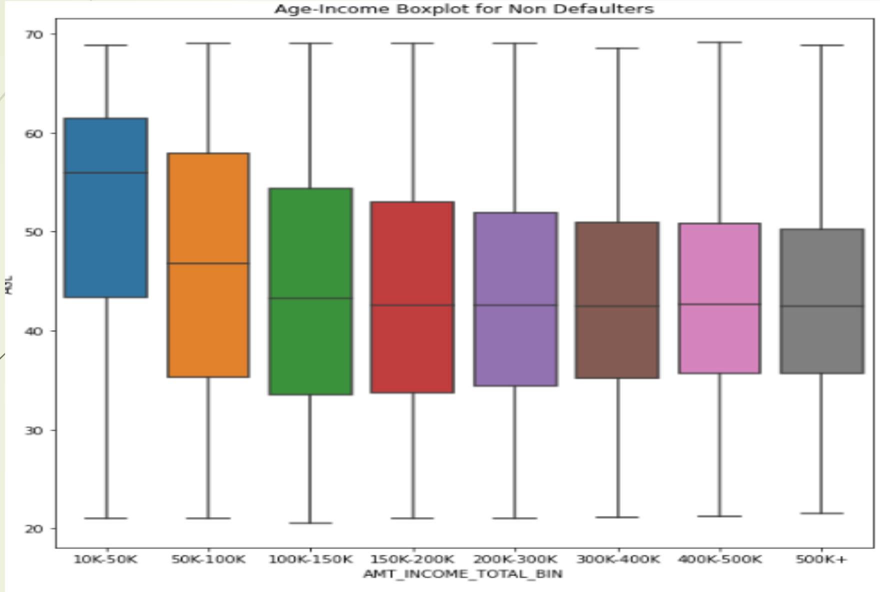


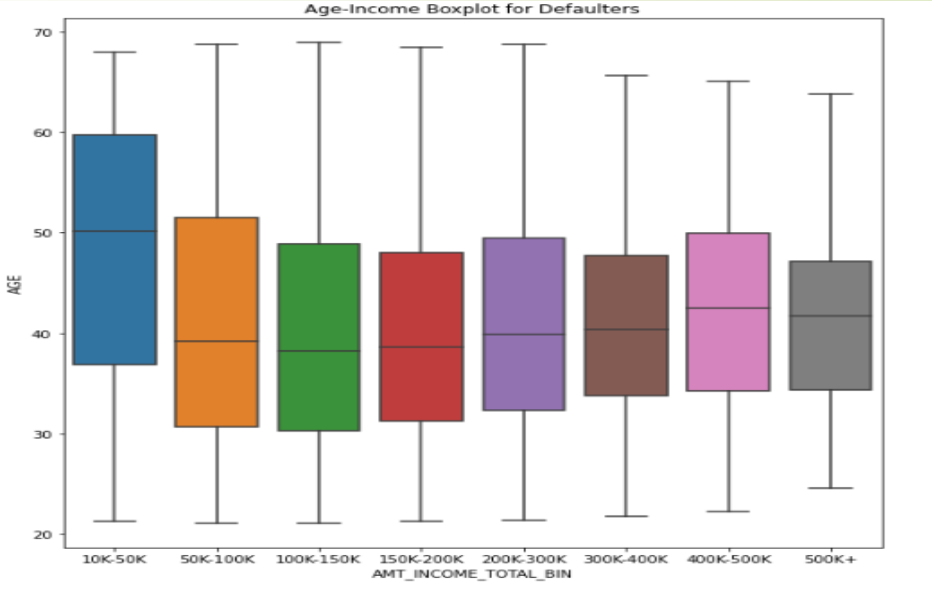


Summary:

- Defaulter/Non-Defaulter Applicants for whom the loan amount was credited between 400K-1M+ majority have family members between 2-3.
- For Defaulter/Non-Defaulter applicants for whom the loan amount was credited < 400K seemed to have majority family size between 1-3 members
- From Income perspective irrespective of the income except < 50K income category, majority defaulters are having 2-3 family members

Age, Income and Target

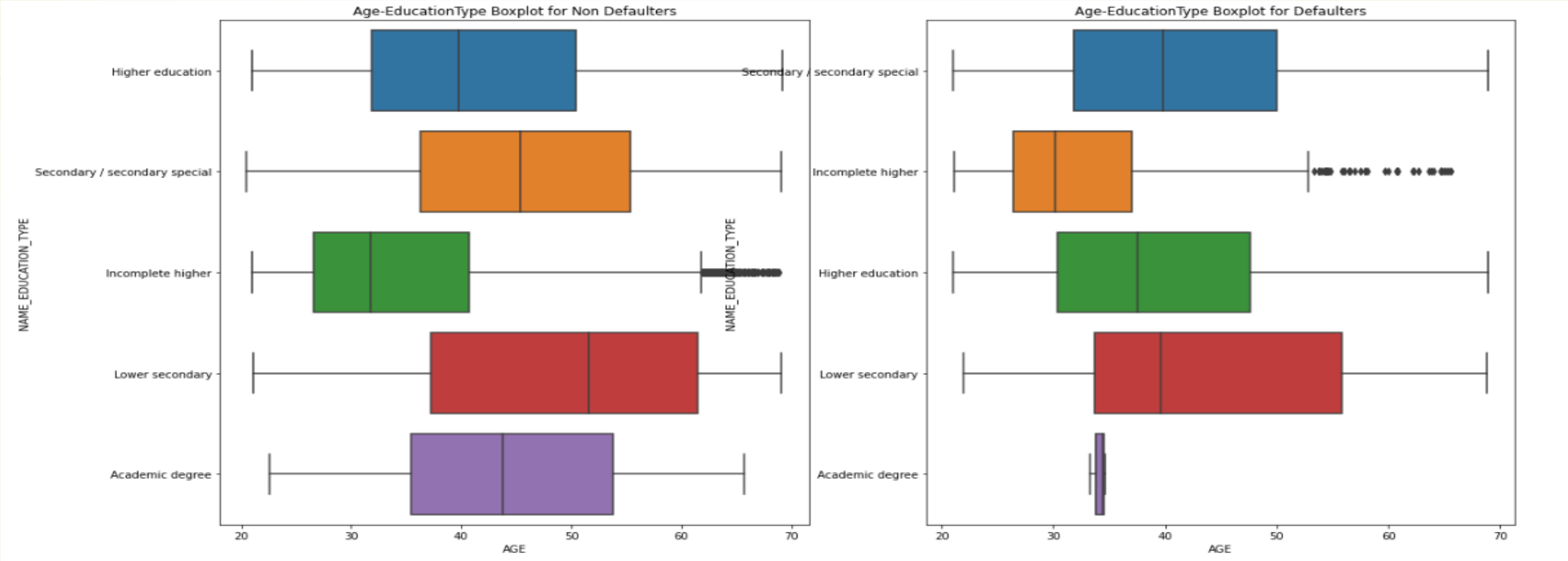




Summary:

- Among the defaulters, across all income group > 50K, 30-50 years applicants with (40 years as median) are most likely to default with some exceptions. For income range < 50K the median age to default jumps from 40 years to 50 years and a range of 38-60 years.
- A similar trend is observed among the non defaulters , < 50K income median age is around 55 years whereas for the other income groups the same drops to a stable range of 43-45 years

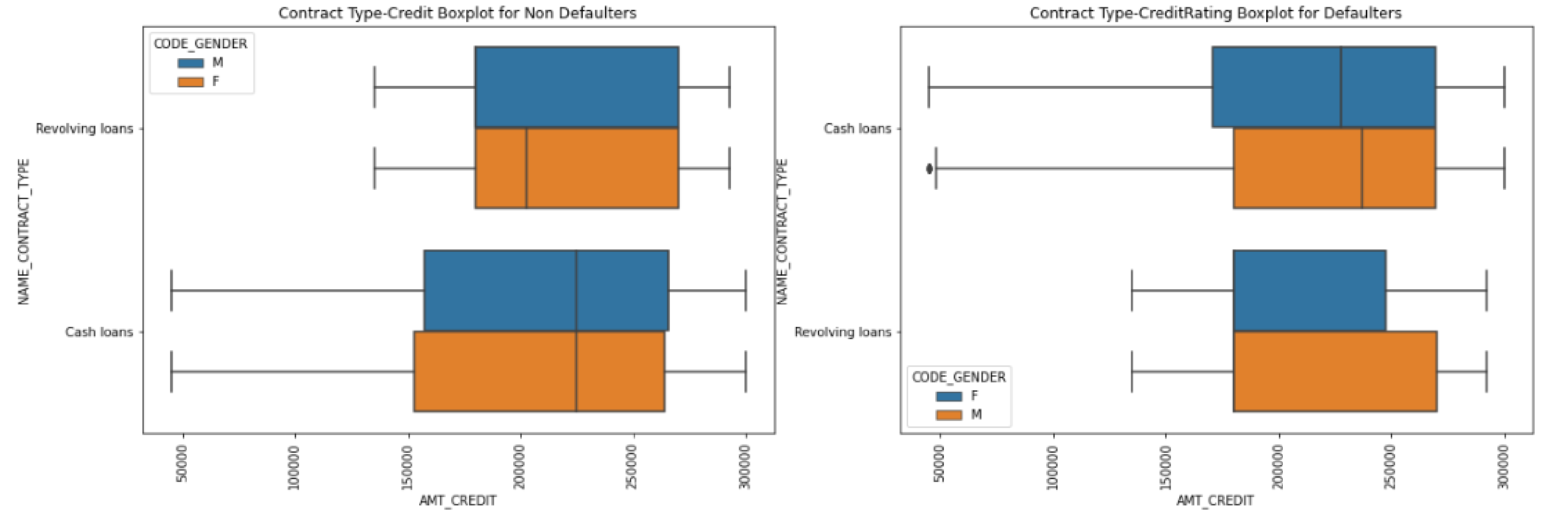
Education Type, Age and Target



Summary:

- Applicants with Academic degree is less likely to default. Median age of such defaulters is around 35 years.
- Applicants with 'Lower Secondary' is most likely to default with a median age of 40 years and IQR between 35-55 years
- 'Incomplete Higher' defaults with a median age of 30 years with an IQR range of 28-35 years.
- For 'Secondary special' and 'Higher education' applicant the median age of defaulter is 35 years with an IQR of 35-50 years

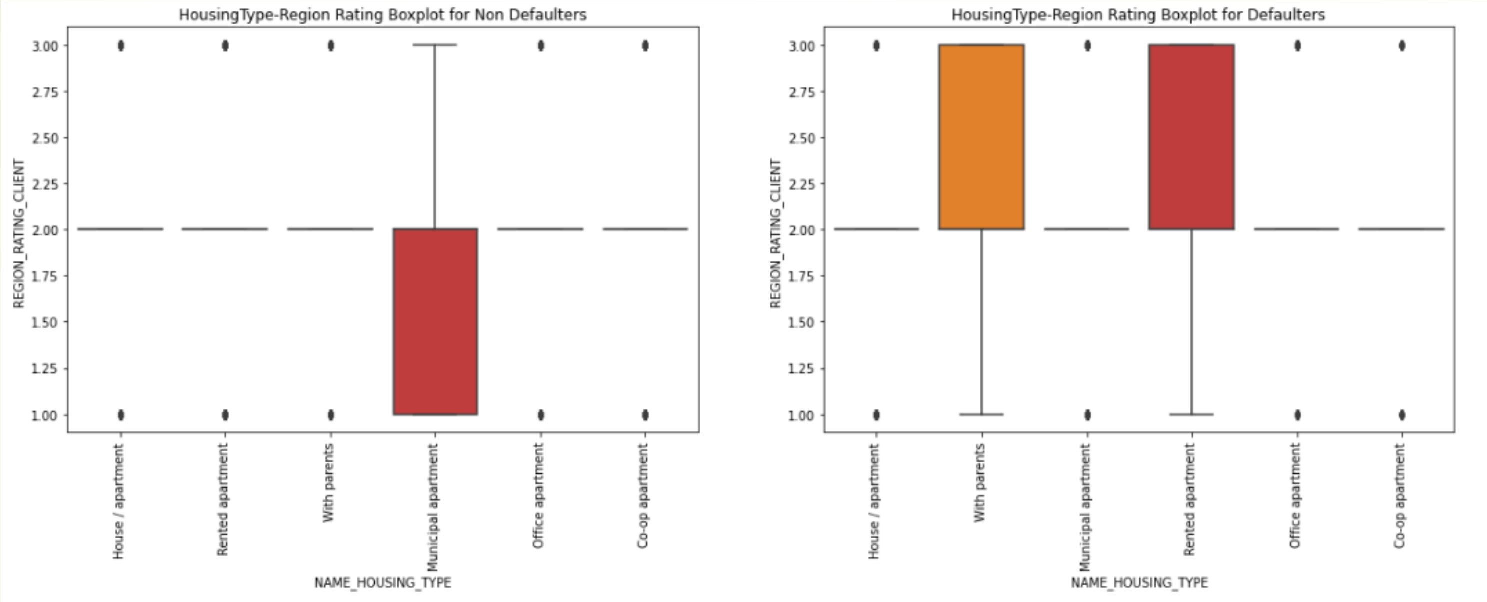
Contract Type, Credit < 300K and Target



Summary:

- For Revolving loans for both Male/Female defaulters AMT_CREDIT amount was between190K – 270K and 190K – 250K respectively. Females seems to be better in paying Revolving loans.
- For Cash loans for both Male/Female defaulters AMT_CREDIT amount is for the same range but Females seemed to have start defaulting from a slightly lesser AMT_CREDIT. Males are better in paying cash loans

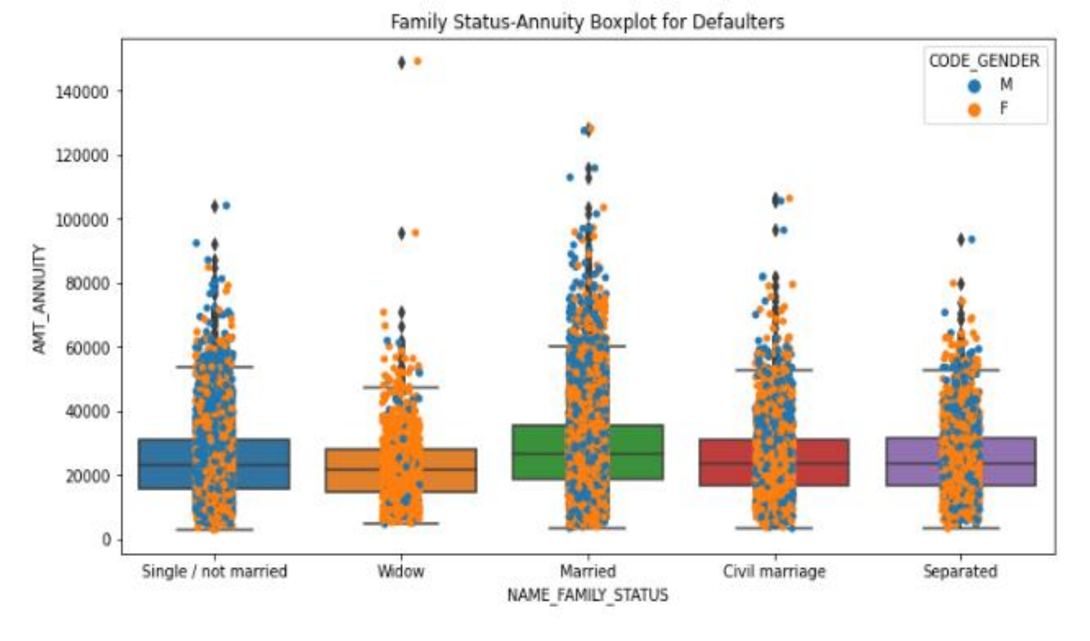
Housing Type, Region Rating and Target



Summary:

- Applicants with a lower REGION_RATING staying in Municipal apartments are good in paying the loans.
- Applicants having higher REGION_RATING but staying ‘with parents’ or ‘Rented apartments’ are most likely to default.

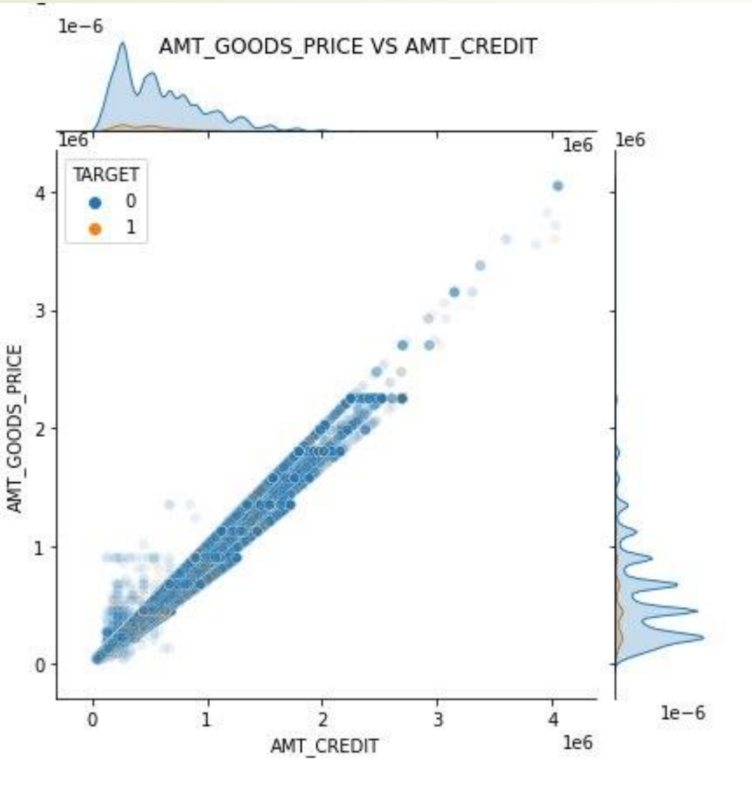
Family Status, Gender, Annuity and Target



Summary:

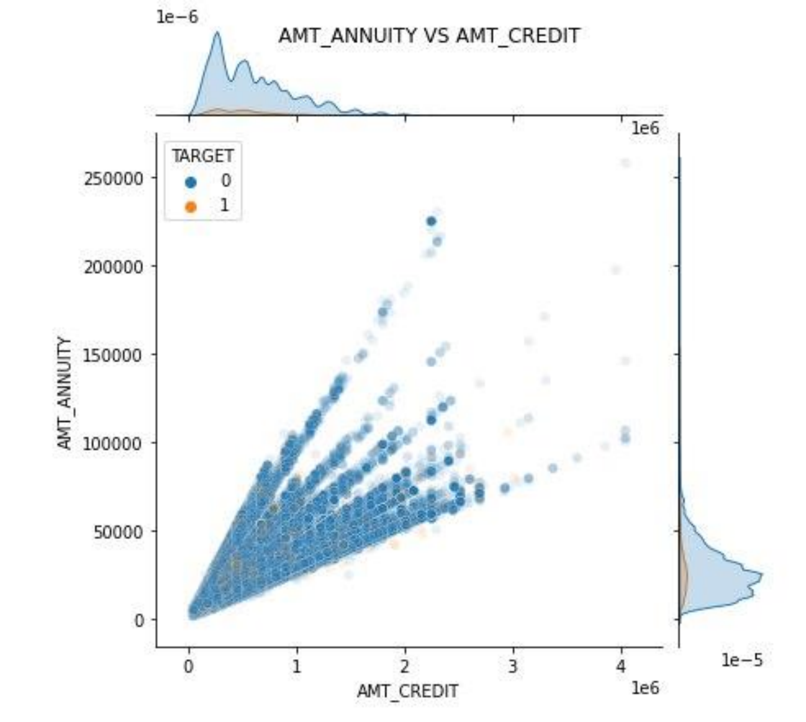
- Widow and Separated Females are more likely to default

Amount Credited Vs Goods Price Vs Annuity



Summary:

- Amount Credited was in line with the goods price with some exceptions. But it seems whatever the Goods price is the same amount was credited.



- Based on some criteria the Annuity amount varies with the credited amount. The parameter is unknown and needs to be further explored.

Top 10 Correlation

TOP 10 Correlations (Non-Defaulter)

	Variable1	Variable2	CORRELATION	CORR_ABS
789	FLAG_EMP_PHONE	DAYS_EMPLOYED	-0.999758	0.999758
1950	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508	0.998508
364	AMT_GOODS_PRICE	AMT_CREDIT	0.987250	0.987250
1219	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149	0.950149
1082	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571	0.878571
1463	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861	0.861861
2011	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859332	0.859332
1646	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830381	0.830381
365	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686	0.776686
304	AMT_ANNUITY	AMT_CREDIT	0.771309	0.771309

TOP 10 Correlations (Defaulters)

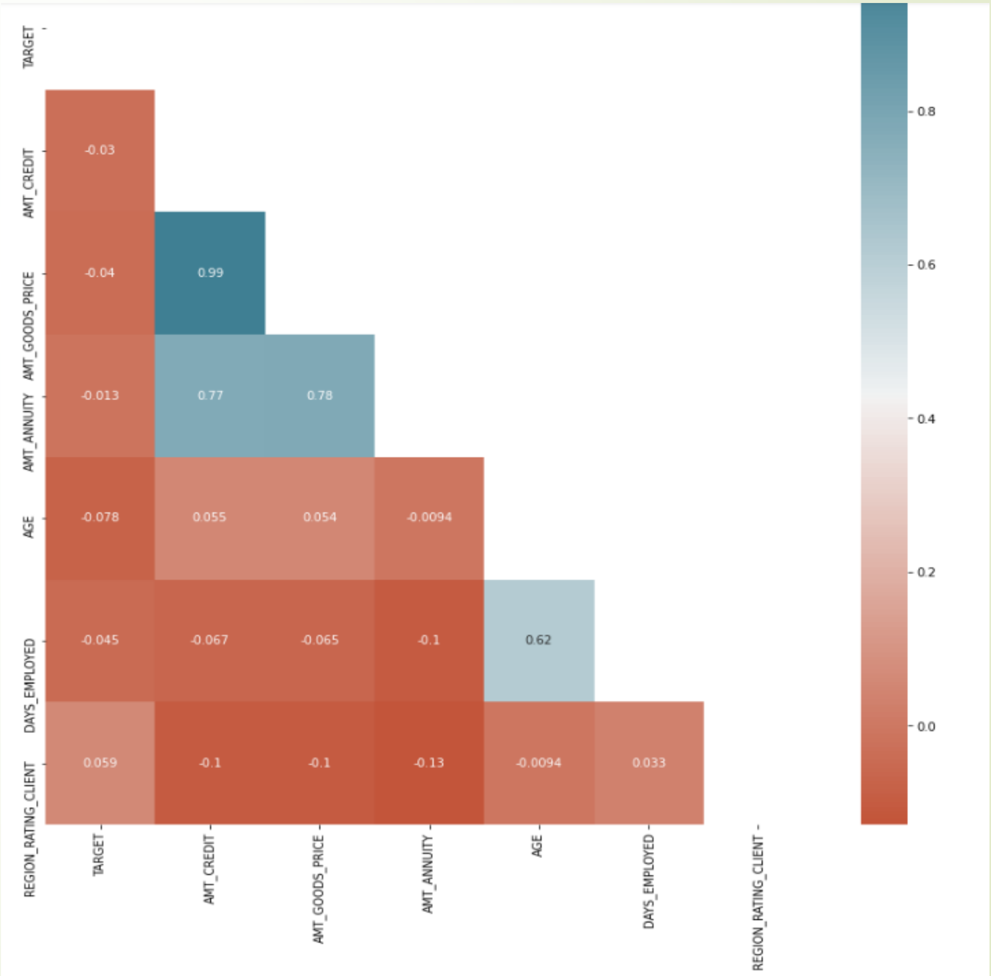
	Variable1	Variable2	CORRELATION	CORR_ABS
789	FLAG_EMP_PHONE	DAYS_EMPLOYED	-0.999702	0.999702
1950	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269	0.998269
364	AMT_GOODS_PRICE	AMT_CREDIT	0.983103	0.983103
1219	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637	0.956637
1082	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484	0.885484
2011	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868994	0.868994
1463	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885	0.847885
1646	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540	0.778540
365	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699	0.752699
304	AMT_ANNUITY	AMT_CREDIT	0.752195	0.752195

CORRELATION (continued..)

Summary:

- There is a high correlation between AMT_GOODS_PRICE and AMT_CREDIT, DAYS_BIRTH and DAYS_EMPLOYED

Correlation for LOAN AMOUNTS / DAYS_BIRTH & DAYS_EMPLOYED

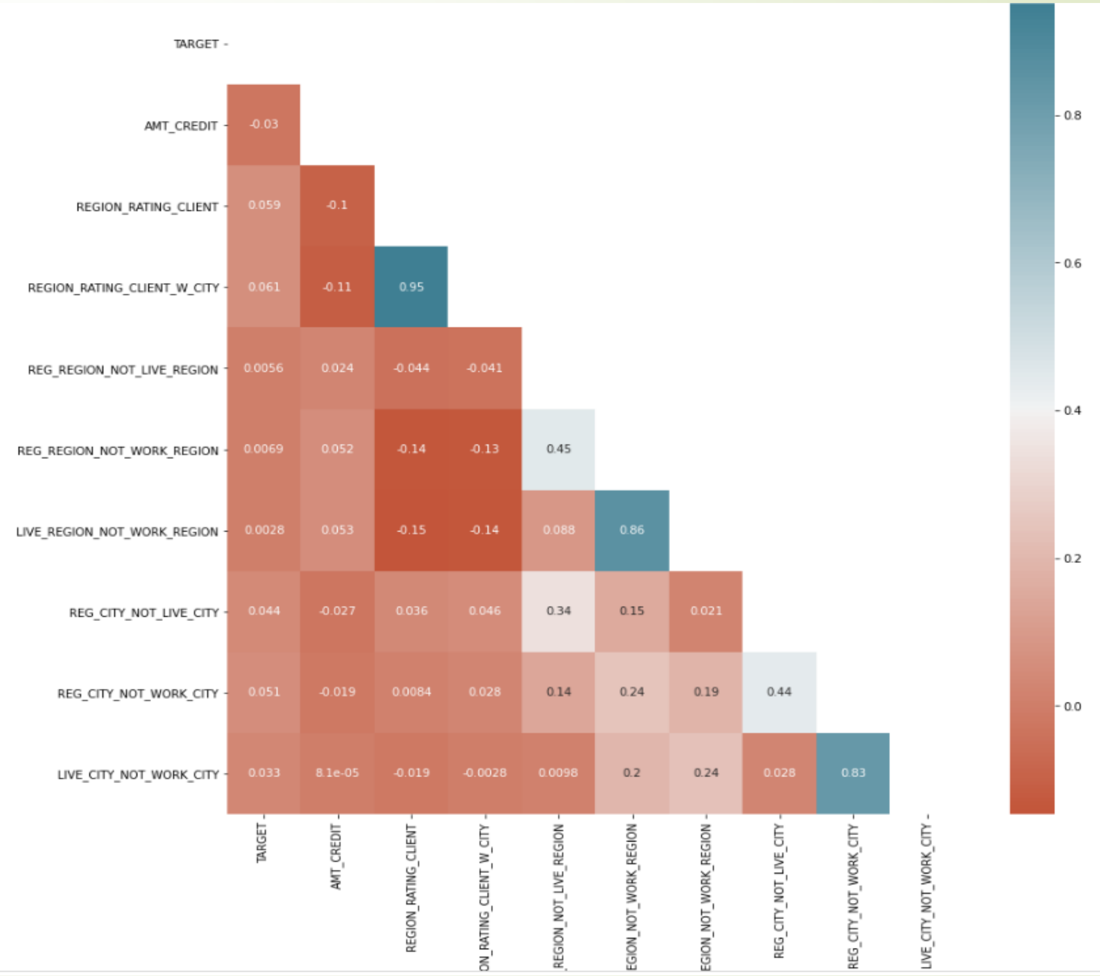


52 CORRELATION (continued..)

Summary:

- There is a high correlation between
  - REGION_RATING_CLIENT and REGION_RATING_CLIENT_W_CITY
  - REG_CITY_NOT_WORK_CITY and LIVE_CITY_NOT_WORK_CITY
  - REG_REGION_NOT_WORK_REGION and LIVE_REGION_NOT_WORK_REGION

Correlation for more variables



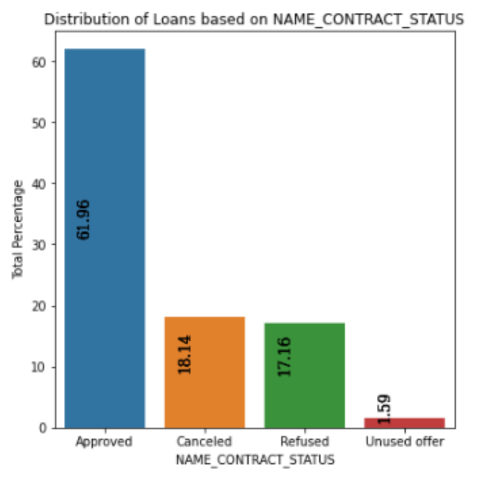
Applicant Historical Loan Status - Analysis

The current application is now merged with the previous application dataset to find out the previous contract status.

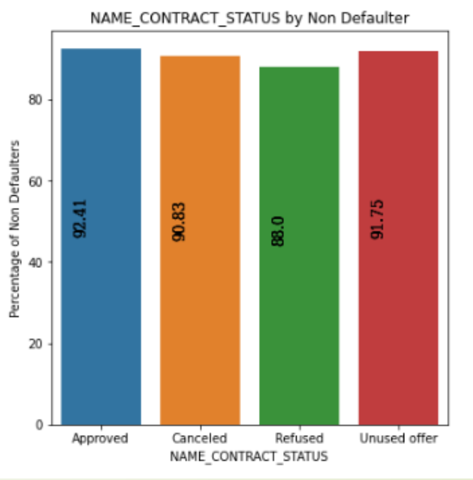
Summary:

- Majority Current Applications previous loan applications was Approved(61.96%).
- If segmented by Defaulters , it seems that majority Applicants who defaulted their previous application was REFUSED.

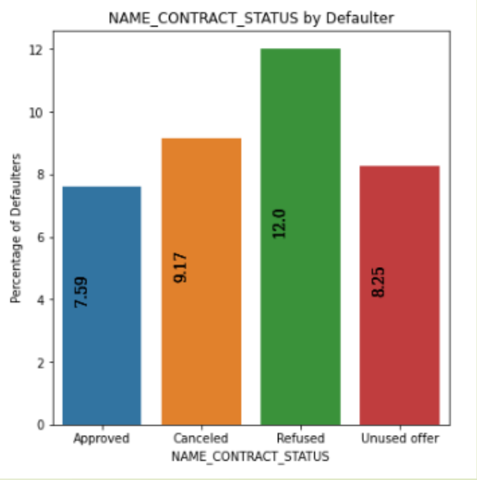
Applicants Historical Loan applications



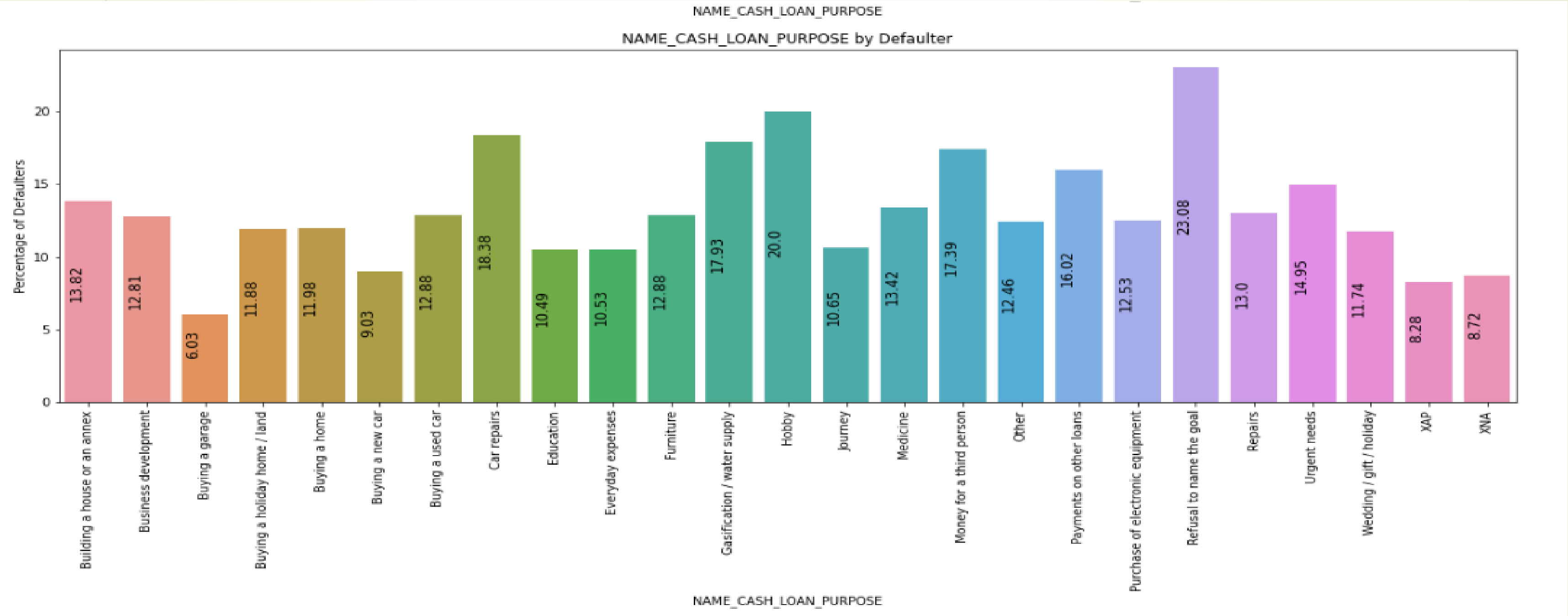
By Non-Defaulter



By Defaulter



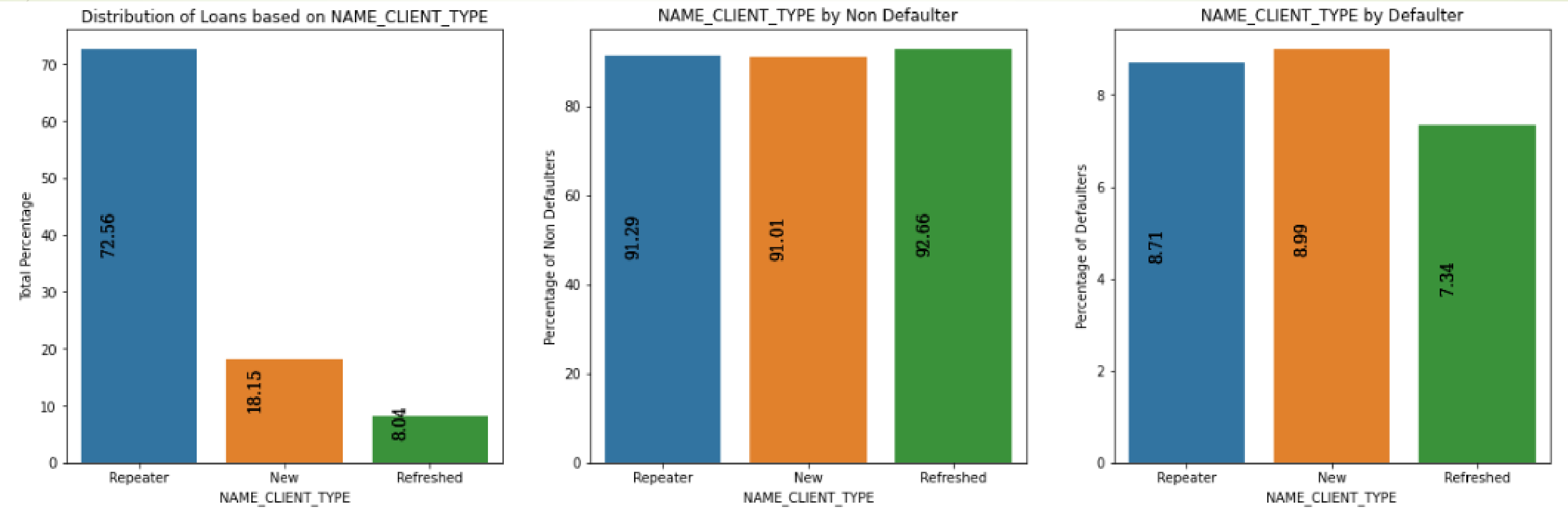
Applicant Historical Loan Status – Analysis (continued..)



Summary:

- Regarding the Cash loans applicant who have taken a loan without a real purpose like *'Refusal to name the goal', 'Hobby' etc.* is more likely to default.

Client Type – Analysis



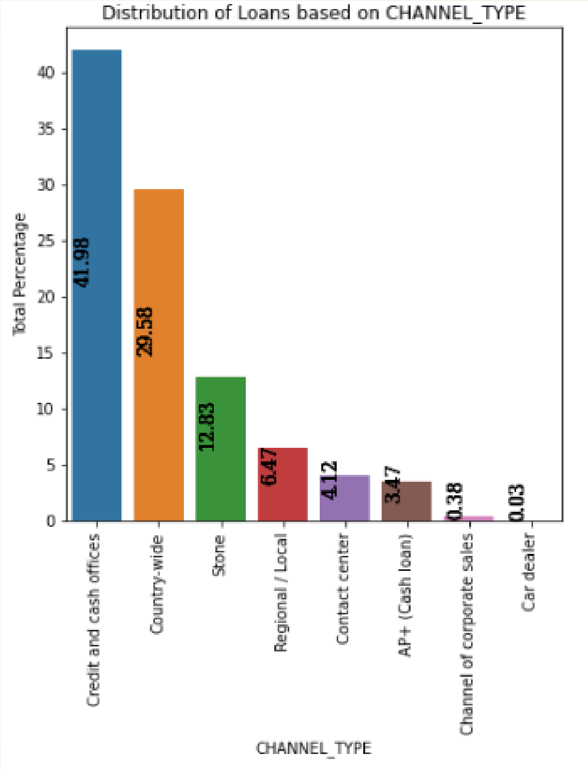
Summary:

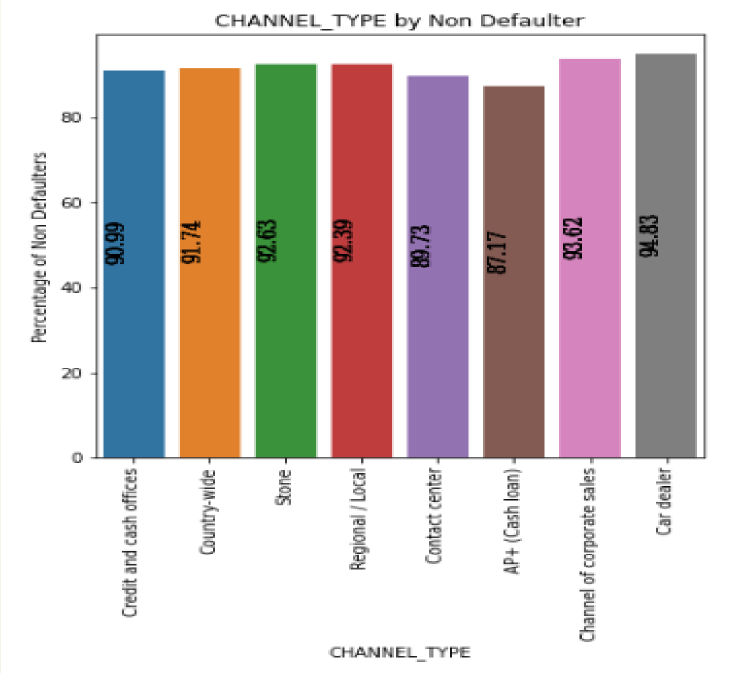
- Majority applications are from 'Repeater' (72.56%).
- Though 'NEW' applicants have defaulted the most. But please note that since 'Repeater' volume is significantly higher than 'New' applicants Repeater will the topmost defaulter.

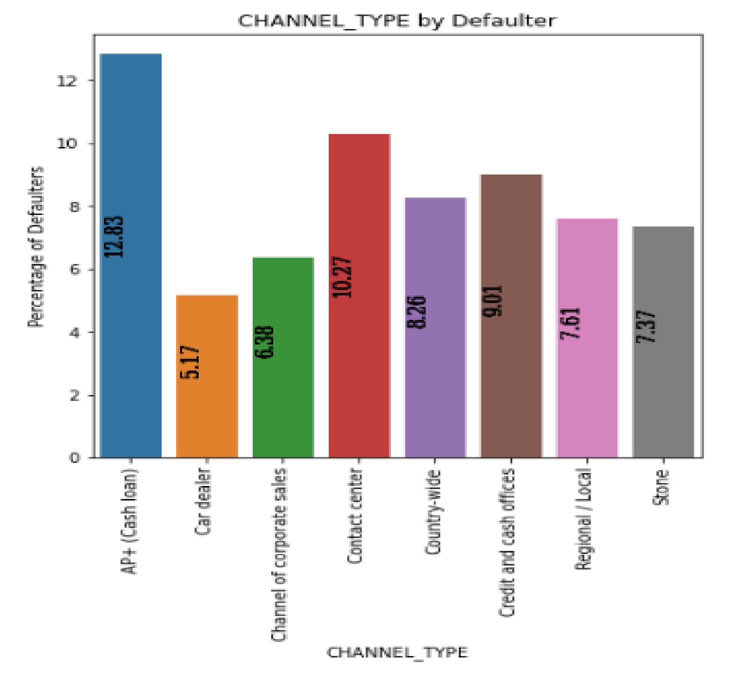
Channel Type – Analysis

Summary:

- Majority applications was channeled via Credit and Cash offices.
- Whereas the loan applications coming from AP+(Cash loans) have the most defaulters from % perspective. Due to the 'Credit and Cash' office volumes it will have the highest defaulter.
- Car dealers are less likely to default



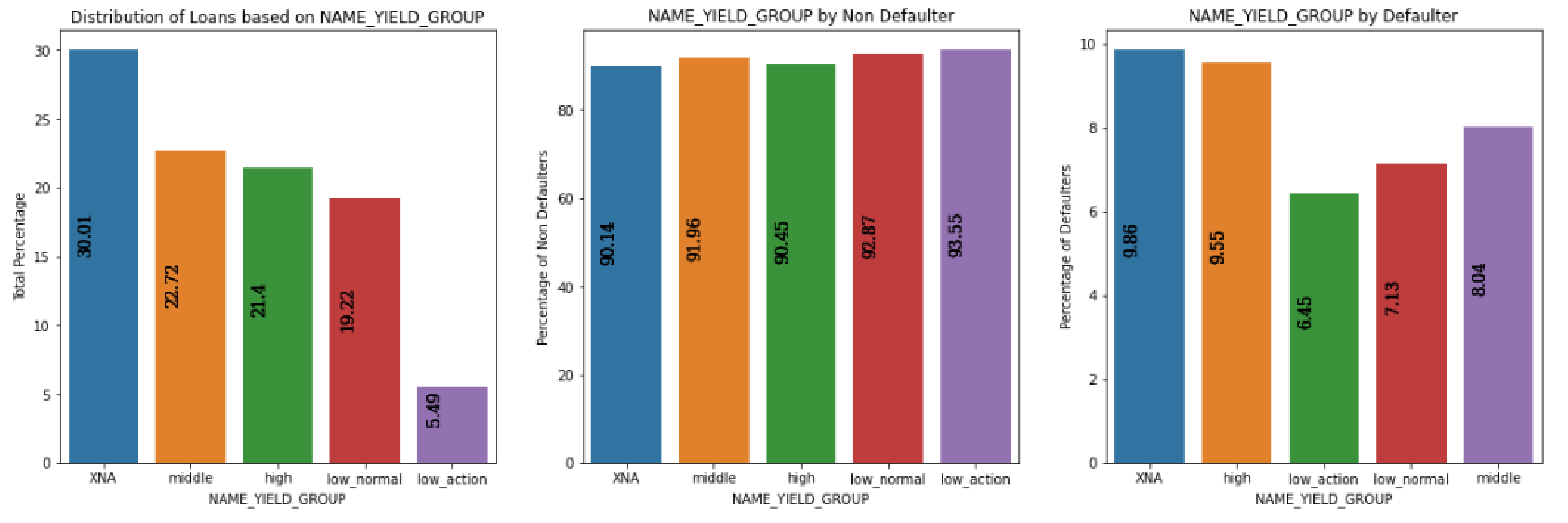




Yield Group – Analysis

Summary:

- Clients with high yield group is more likely to default compared to low_action group
- Clients who do have a yield group, set as 'XNA' also likely to default. Not clear why 'XNA' was set against NAME_YIELD_GROUP. Maybe its not known or there is an error with the dat. Further analysis required for the same.

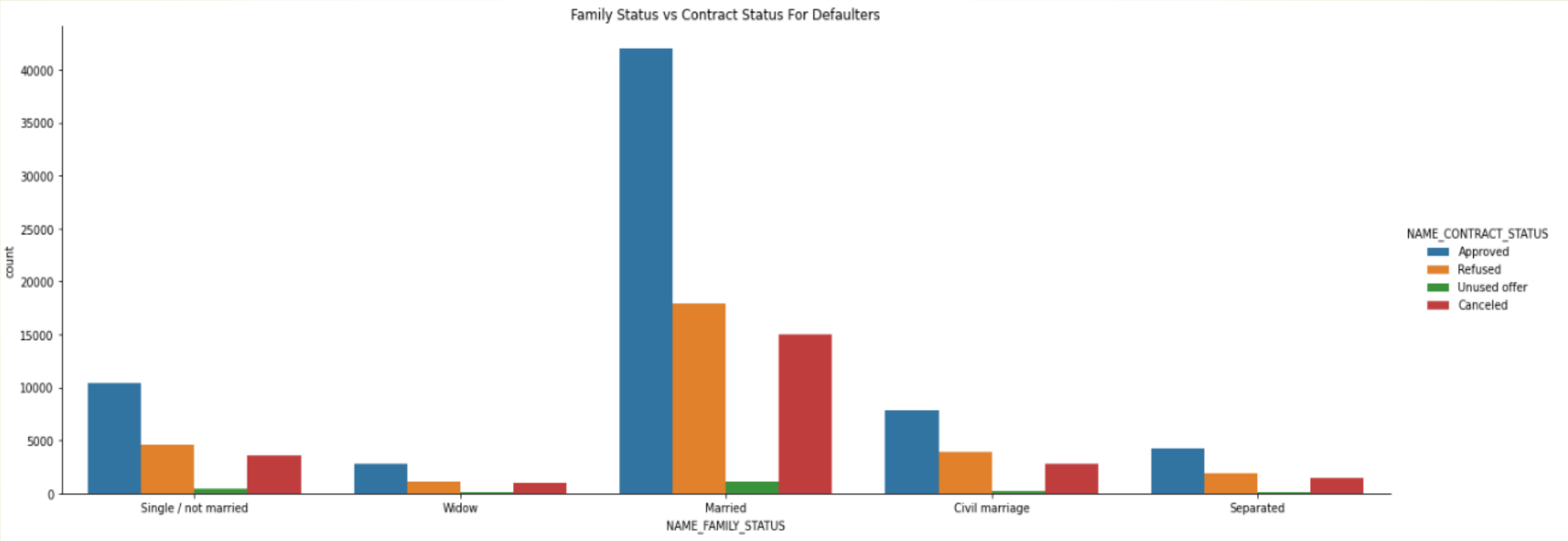


Family Vs Previous Contract Status

Summary:

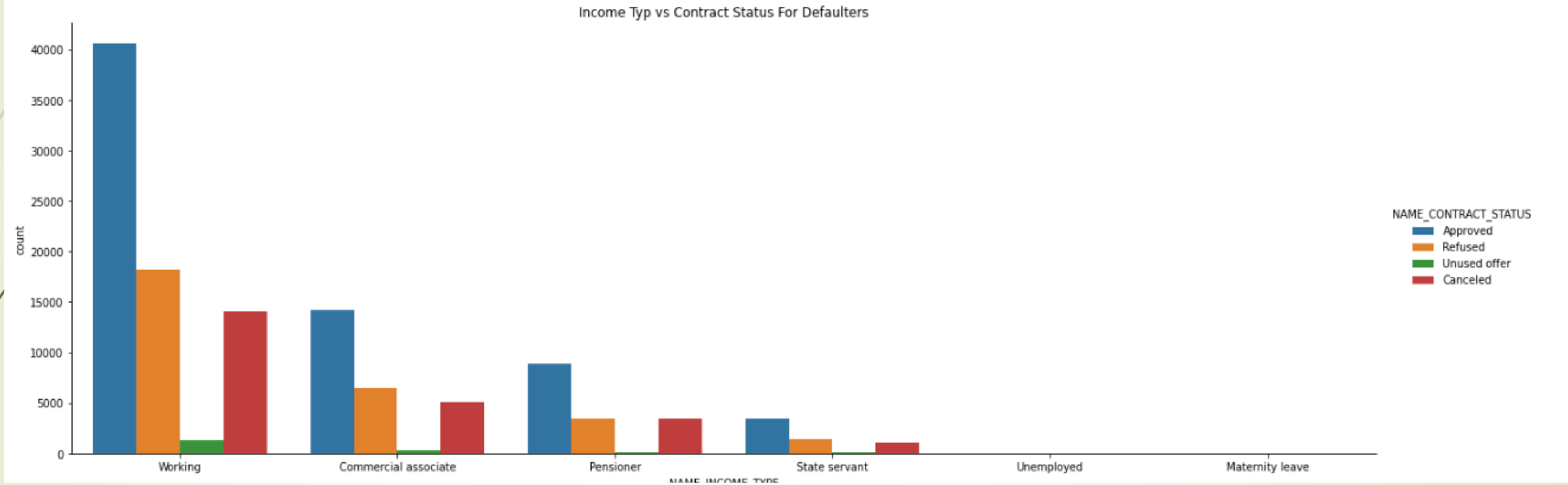
❖ Married applicants are the majors loan applicant category. Based on the graph below even though their previous loans was approved/refused or unused, the category are more likely to default. Having said that please note that 'Married' category is also one of the highest non-defaulters therefore there are other parameters which are making applicant of the category to default.

Family Status Vs Previous Contract Status for Defaulters



Income Type Vs Previous Contract Status

Income Type Vs Previous Contract Status for Defaulters



Summary:

- Similar to the Married Applicants, applicants having Working are the highest loan applicants and defaulters as possible. It doesn't matter if the previous loan application was approved/refused or unused. There are other driving parameters which is making applicant to default.

Conclusion

Summary:

- Bank should give more loans to clients with academic degree as they default less. Clients with Lower secondary educations are more likely to default.
- Low skill laborer's, drivers are more likely to default loan when compared to high skill occupations.
- Applicants who had previous loan applications Refused and Cancelled are more likely to default loans.
- Applicants with lower EXT_SOURCE_2 and EXT_SOURCE_3 (Rating) scores are likely to default.
- Applicant staying in Rented Apartment / Staying with Parents are more likely to default. Also, since major loan applicants resides in House/Apartment a lower default rate would make a good number of defaulters in the category.
- Clients from Low populated regions are more likely to default more compared to Highly populated regions.
- Cash loans applicant who have taken a loan previously with purpose 'Refusal to name the goal', 'Hobby' etc. are more likely to default.
- Applicants who owns a house is are more likely to default. It may be because of additional liabilities they have . Therefore, other parameters like needs to be checked for the applicant.

Conclusion (continued…)

Summary:

- Overall clients > 40% are less likely to default.
- Banks should check the applicant's income type (e.g., Maternity leave / Unemployed) who are more likely to default. Also, since Commercial Associates reflect in the Top 3 defaulter / Loan applicant list hence it will have a larger impact if 7.48% Commercial associate defaults.
- Applicants who have just started work are most likely to default. Definitely other variables needs to be taken for a decision.
- Higher Region/City Rating applicants are less likely to default.
- Applicants who are Single or have a Marital status as Civil Marriage is more likely to default but since Married applicants apply the most even 7.65% defaulters out of the category will make a good number of defaulters.
- Applicants who are 'New' defaults the most. It may be because the bank didn't have a credit history for the applicant. Repeaters are the next set of defaulters.