BANK LOAN CASE STUDY

Trainity Project Report

Rohit kumar
rohitk.ug20.ce@nitp**.ac.in**

DESCRIPTION

This case study aims to idea of applying EDA in areal business scenario. In this case study, apart from applying the techniques it also focus on basic application of risk analytics in banking and financial services and describe how data is used to minimize the risk of losing money while lending to customers.

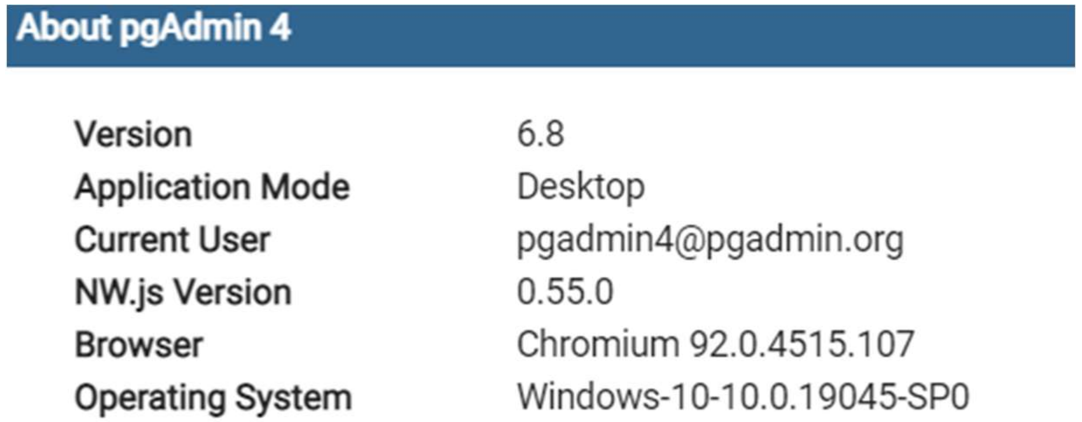
About Dataset

- **‘application_data.csv’** contains all the information of the client at the time of application this data is about whether an applicant has payment difficulties.
- **‘previous_application.csv’** contains information about the applicant’s previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
- **‘application_data.csv’** dataset have 122 column and 307511 rows and **‘previous_application.csv’** have 37 column & 1670214 rows
- **‘application_data.csv’** have 110 numerical column (including Boolean and district type) 16 string type column
- **‘previous_application.csv** have 21 numerical column (including Boolean and district type) 16 string type column

This original Dataset is very huge for implementation of Analysis process in MS-Excel. Hence here I will use sample dataset for further analysis

TECH STACK USED

- **PostgreSQL PG Admin 4 version 6.8** was only using extract Sample from Original Dataset



About pgAdmin 4

Version	6.8
Application Mode	Desktop
Current User	pgadmin4@pgadmin.org
NW.js Version	0.55.0
Browser	Chromium 92.0.4515.107
Operating System	Windows-10-10.0.19045-SP0

- '**Microsoft Excel 2016**' was used to perform Analysis
- **'MS Power Point Presentation 2016'** was used to prepare to report.

Click below to view excel sheet containing steps & solution

- SOLUTION 1
- SOLUTION 2

APPROACH

- To perform Analysis on data records I begin with choosing right column and extracting it for further cleaning process
- After cleaning, filtering and sorting the dataset, it time to process the data,
- Here I use excel function such as advance filter, Text to column, mathematical function, Pivot Table, Data analysis Ads in function etc. and also apply different combinations and visualize it to under stand it better.
- Expect cleaning process all analysis is done by using Pivot table and basic excel function

FINDINGS

PROBLEM STATEMENT

The objective of this case study is to identify patterns that indicate that a borrower is having difficulty repaying a loan, which can lead to such actions as denying the loan, reducing the amount of the loan, and lending (to high-risk borrowers) at a higher interest rate. Consumers who are capable of repaying the loan will not be rejected, and the number of defaulters will also be reduced.

It is imperative for banks to understand the factors(or variables) responsible for loan defaults, i.e. the factors that are proven to be strong indicators of default. This knowledge can be used to assess the bank’s portfolio and risk

DATA CLEANING & MANIPULATION

STEPS FOR APPLICATION DATASET

1. Counted missing values by using “COUNTBLANK” & calculated percentage of missing values in dataset

2. There 49 column which have more that 35 % missing values
3. Thereafter I check some correlation insignificant column of with **TARGET** column by using **Ads in Option DATA ANALYSIS in DATA tab** and found no relation

4. And finally deleted **78 columns** as mentioned beside

5. Then fill the missing data using appropriate method belonging to feature
6. DAYS Birth Change in year format and DAYS_EMPLOYED,DAYS_REGISTRATION, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE changed in month format and change column name

7. There **are 556 wrong value** present in EXP_In_YEARS column **i.e., 1000.7** years in which most of client is above **50 years old** we'll keep this wrong value as it is... as it not relevant to any other column
8. Impute with median as mean has decimals and this is number of requests in column Amount related column and days related column

9. Impute categorical variable **'OCCUPATION_TYPE'** which has higher **null percentage (32.09%)** with a new category **(i.e., UNKNOWN)** as assigning to any existing category might influence the analysis

DATA CLEANING & MANIPULATION

Distrubution of Deleted Column	
Type of col	**No. of col**
missing values more that 35%	49
ext_sources	2
documents col	19
contact details cols	6
week & hours of appli. Proccess	2
TOTAL	**78**

DATA CLEANING & MANIPULATION

STEPS FOR PREVIOUS APPLICATION DATASET

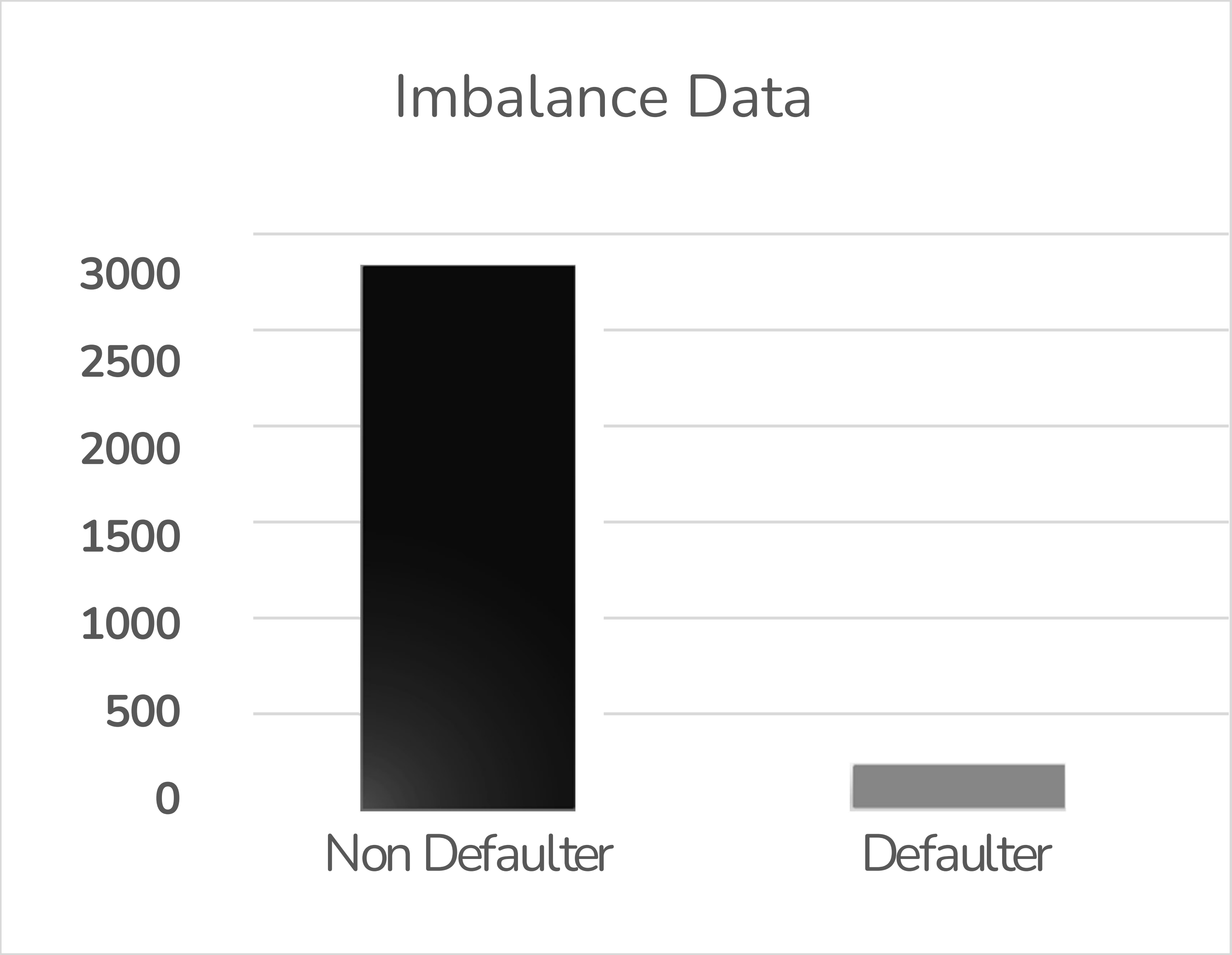
1. Counted & calculated percentage of missing values in dataset
2. Deleted all 5 column which have more that 45 % missing values and 2 column related to Hours & Weekday of application process start
3. Missing values in column DAYS_FIRST_DRAWING, FIRST_DUE, LAST_DUE_1ST_VERSION, LAST_DUE, NFLAG_LAST_APPL_IN_DAY because of refusal of loan application
4. There is around 22 % missing values in AMT_GOODS_PRICE, AMT_ANNUITY, CNT_PAYMENT due to loan application either get CANCEL , REFUSE or unused offer
5. Filled product combination column with MODE (i.e., CASH) using pivot table
6. DAYS_FIRST_DRAWING is not useful column as more that more 55% + filled with same wrong value so it will get drop
7. AMT_Application have same datapoints as AMT_GOODS_PRICE so column AMT_GOODS_PRICE will get drop
8. Add TARGET column from application data to this dataset with VLOOKUP Function formula
   "=VLOOKUP($C2,supporting dataset1!$A$2:$B$3076,2,FALSE)"

DATA IMBALANCE RATIO

client Type	Count of TARGET	pct share
Non Defaulter	2834	**92.16**
Defaulter	241	**7.84**
Grand Total	**3075**	**100**

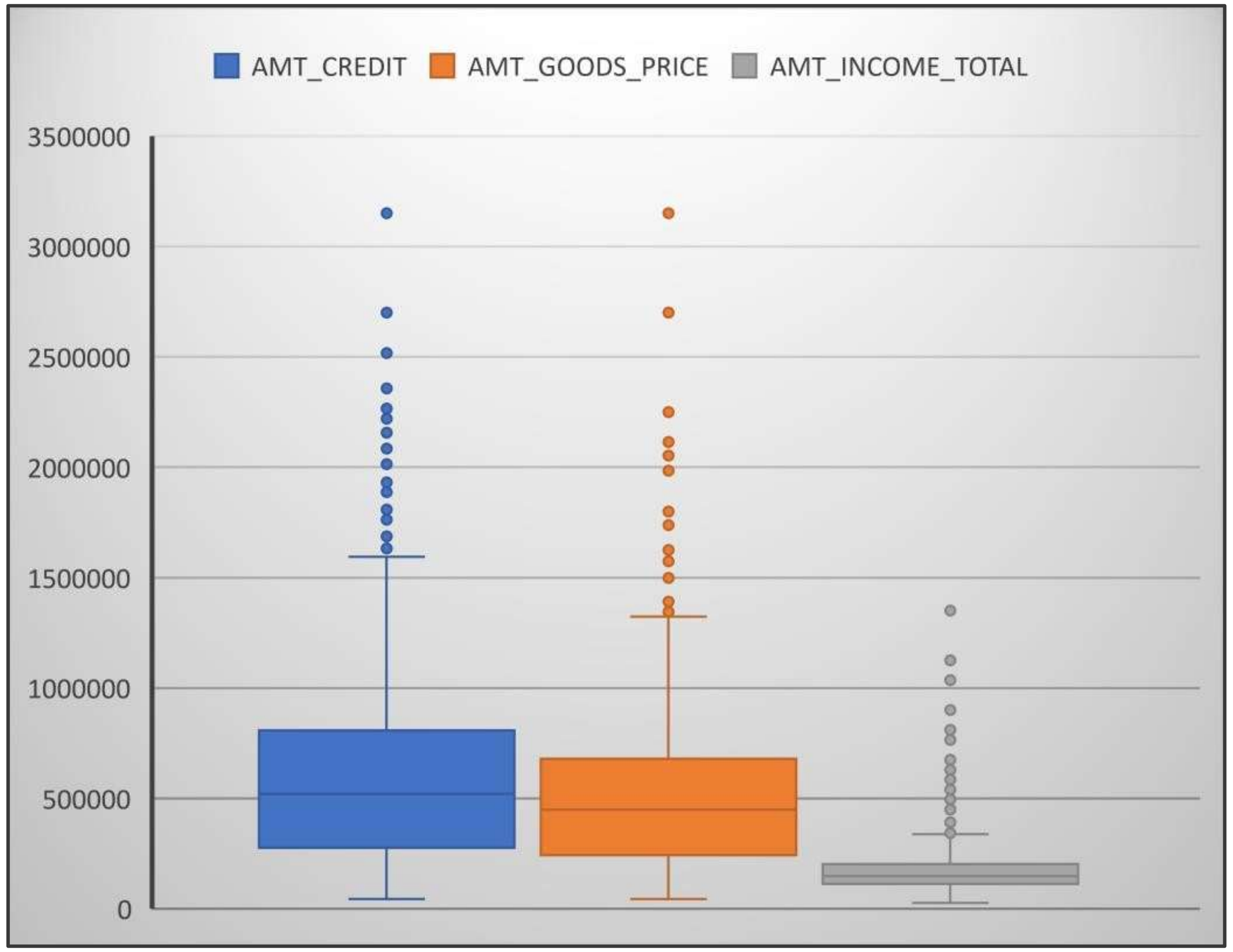
Ratio of imbalance in relative with respect to Non defaulters and Defaulters data is **11.39 : 1**

Percentage of **Non defaulters is 92.16 % & Defaulters is 7.84%**

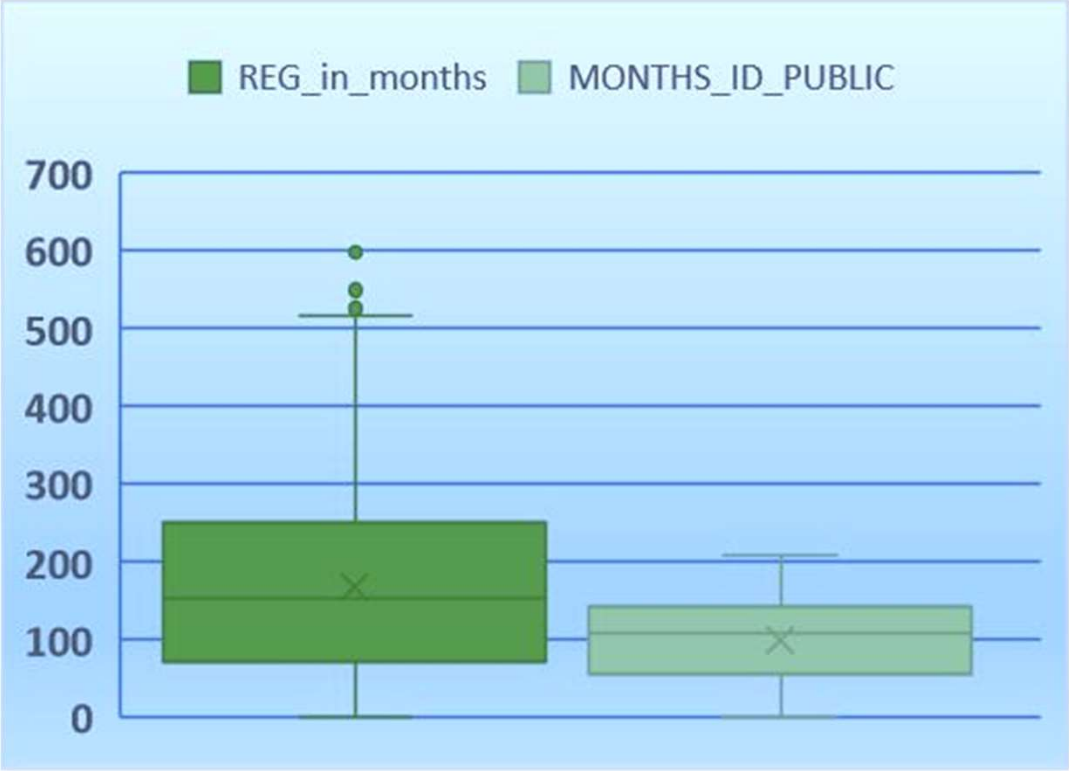


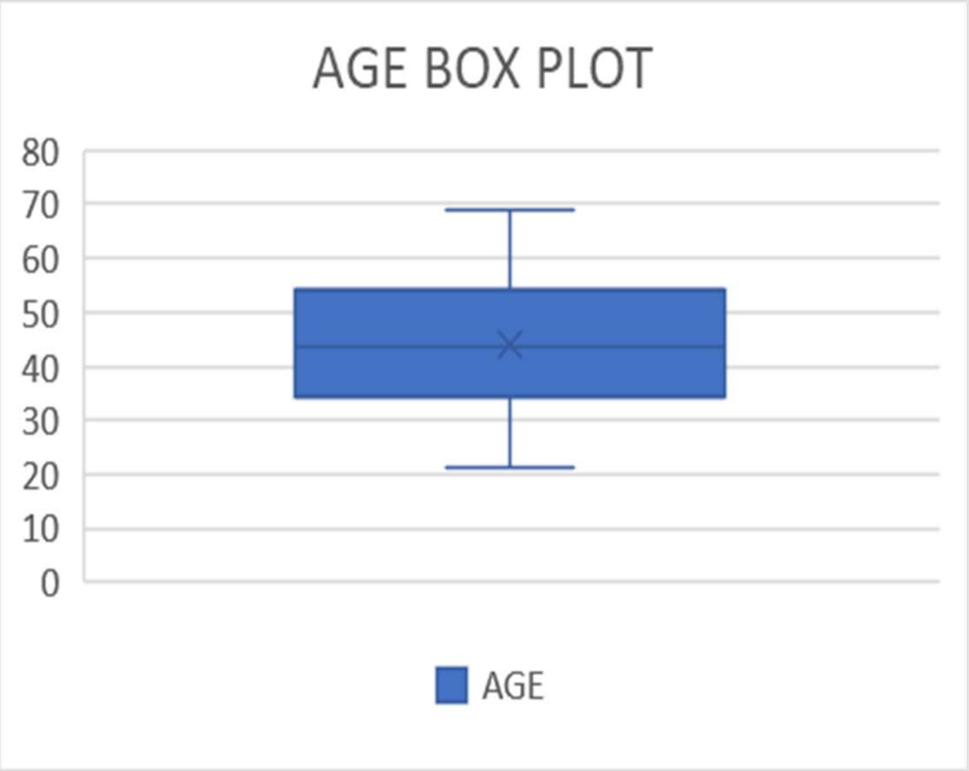
OUTLIERS IN DATA

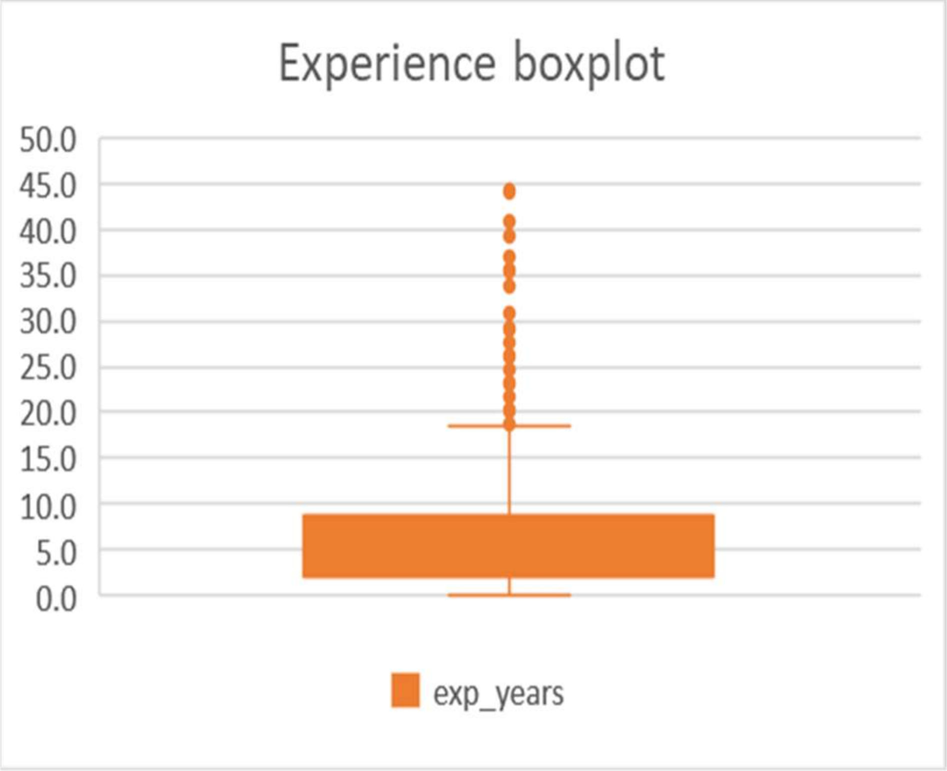
- Amount Related column have Continuous outliers
- Outliers have AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.
- AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have high income when compared to the others.
- All outliers founded place in higher limit



OUTLIERS IN DATA

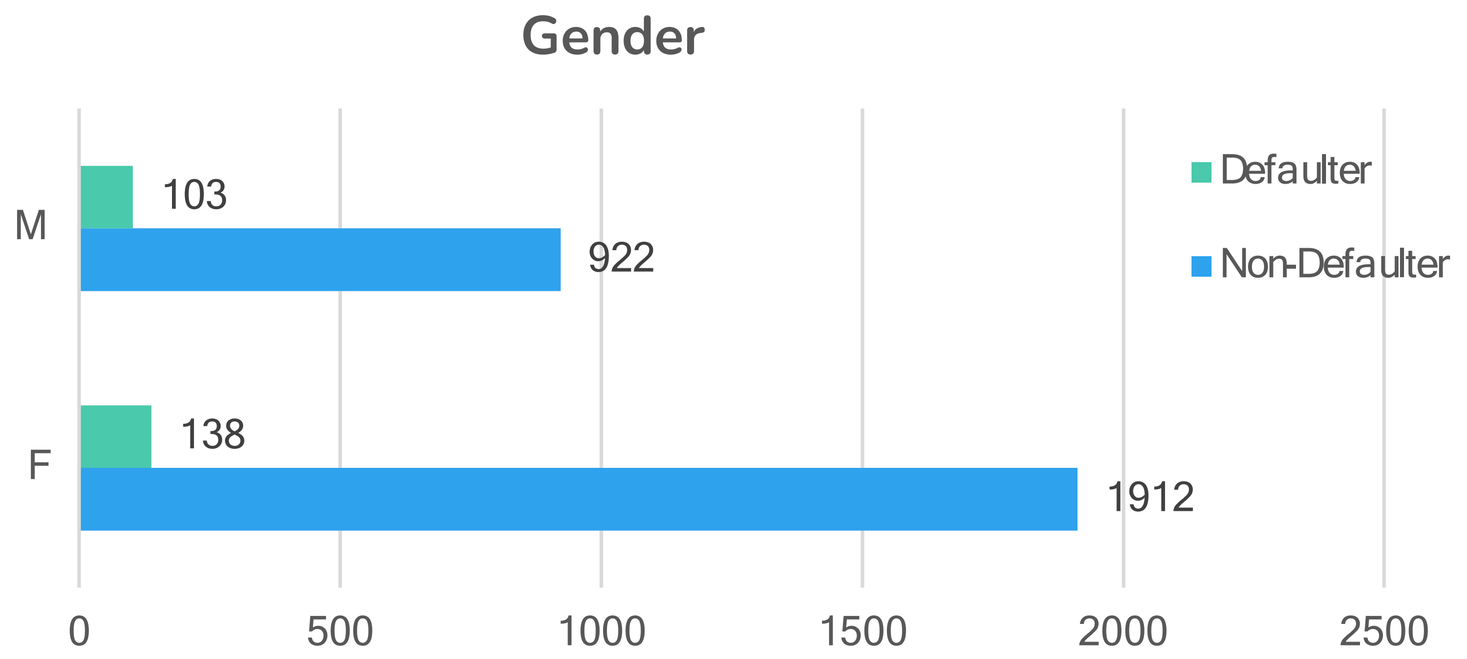


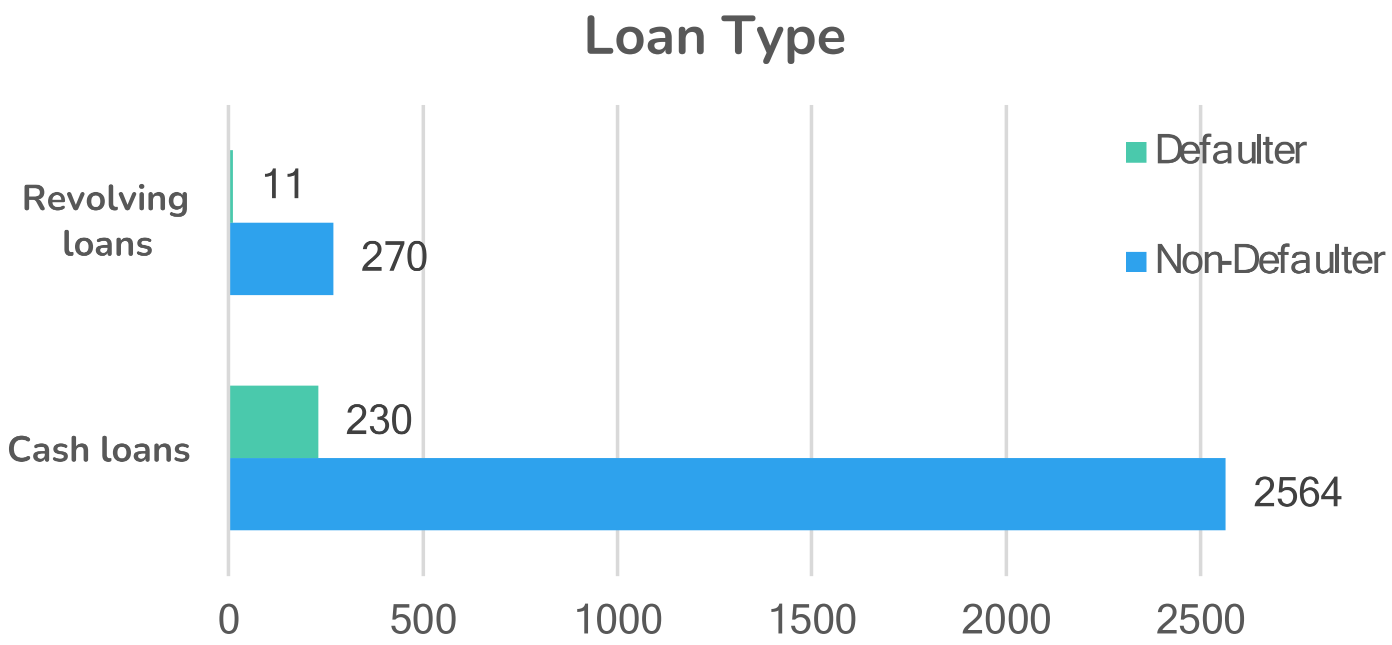




- While compare to other features exp. In Years have lot of outliers present in reg month.
- Most of clients take loan in their career beginning
- Age doesn’t have any outliers

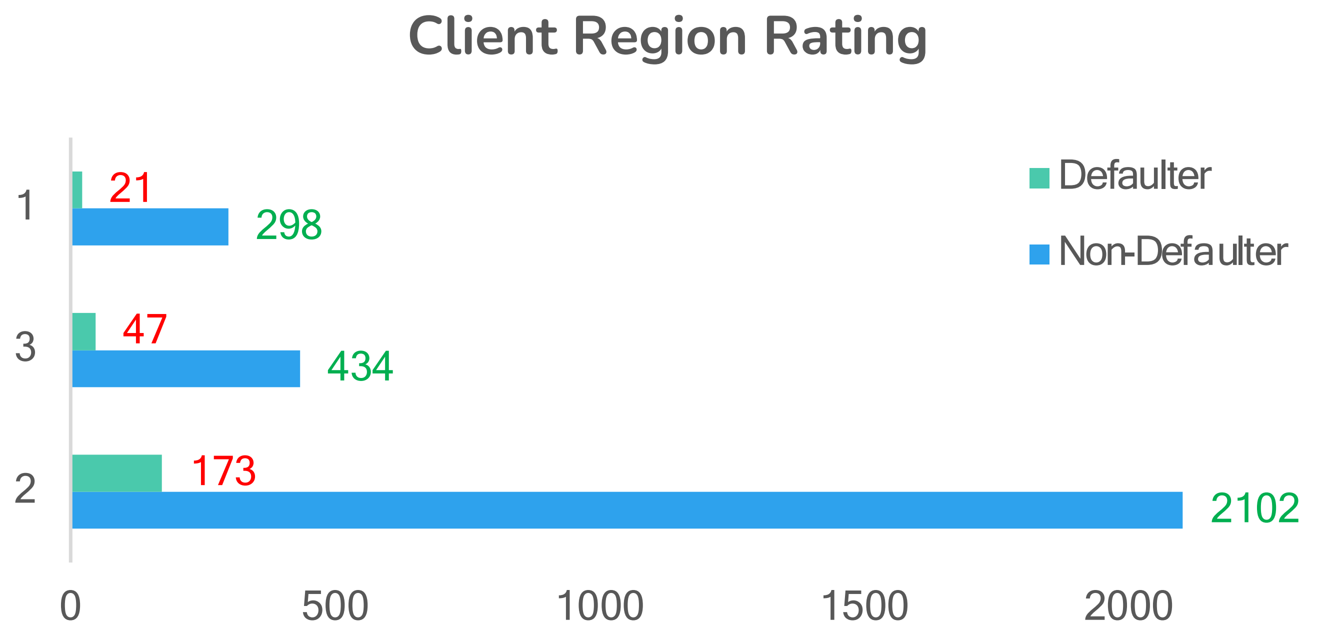
UNIVARIATE & SEGMENTED UNIVARIATE ANALYSIS

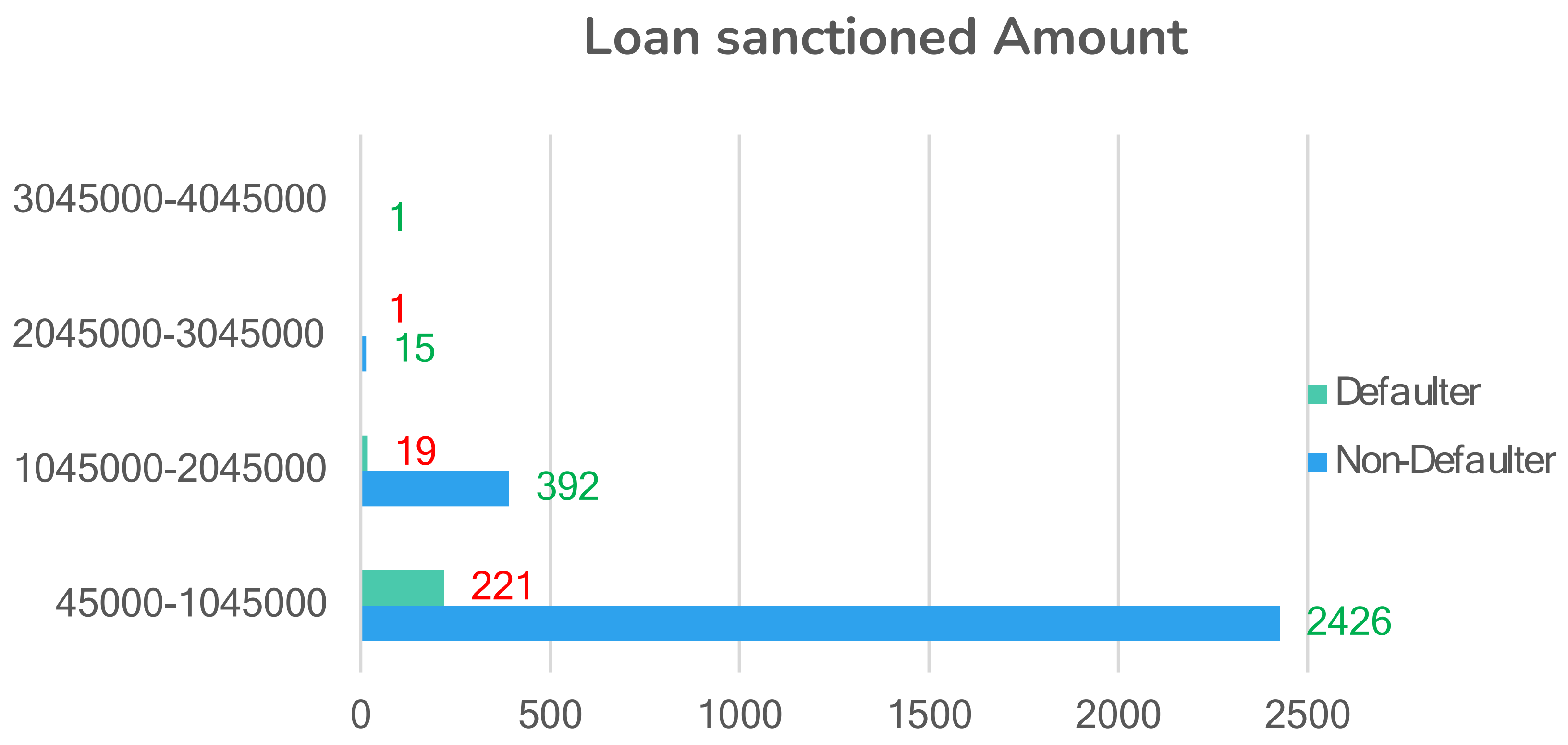




- Male client have higher rate of defaulter as compare to female
- Cash loan have around 10 % of defaulter rate while revolving rate have less i.e. around 4 % of defaulter rate

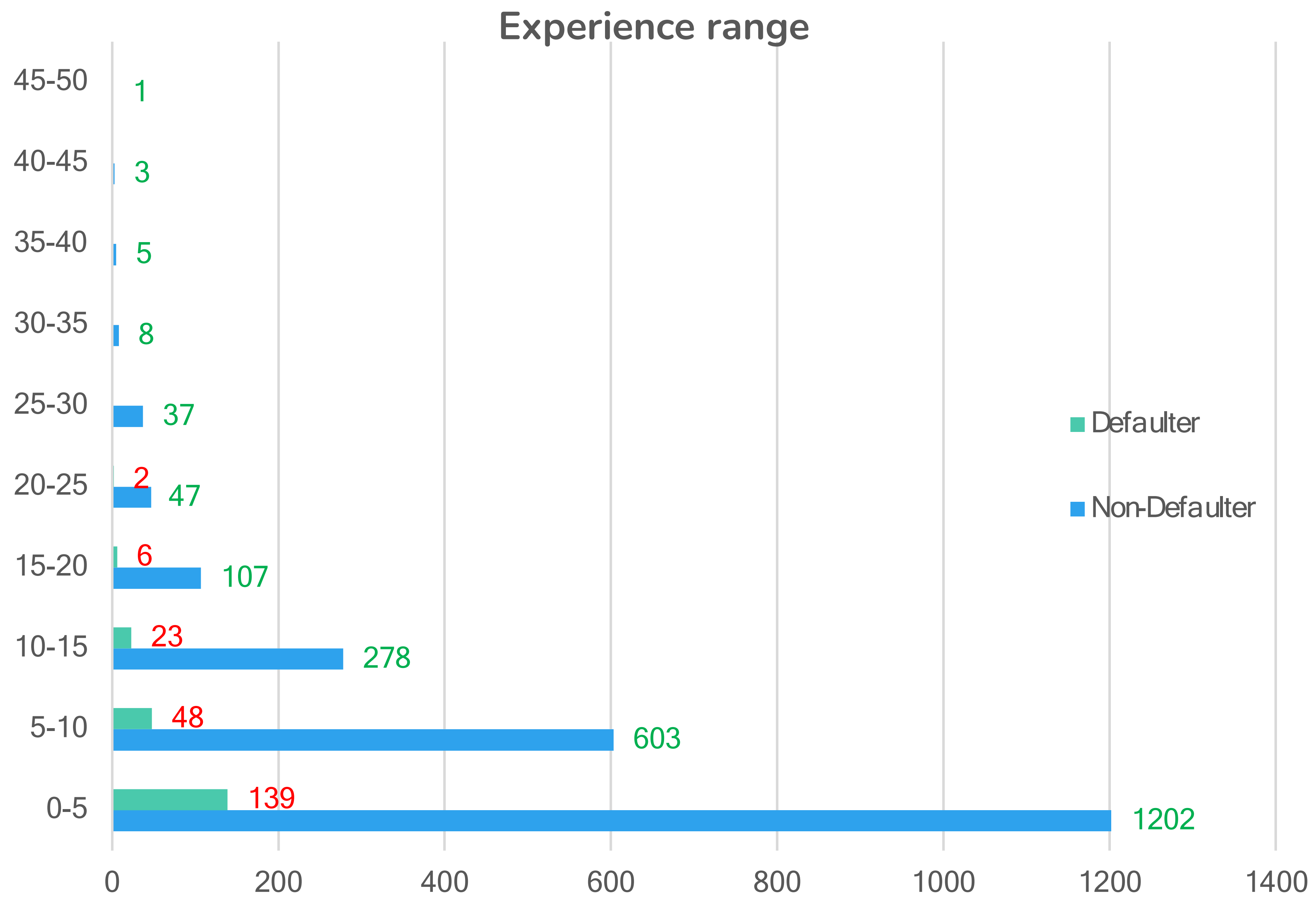
UNIVARIATE & SEGMENTED UNIVARIATE ANALYSIS





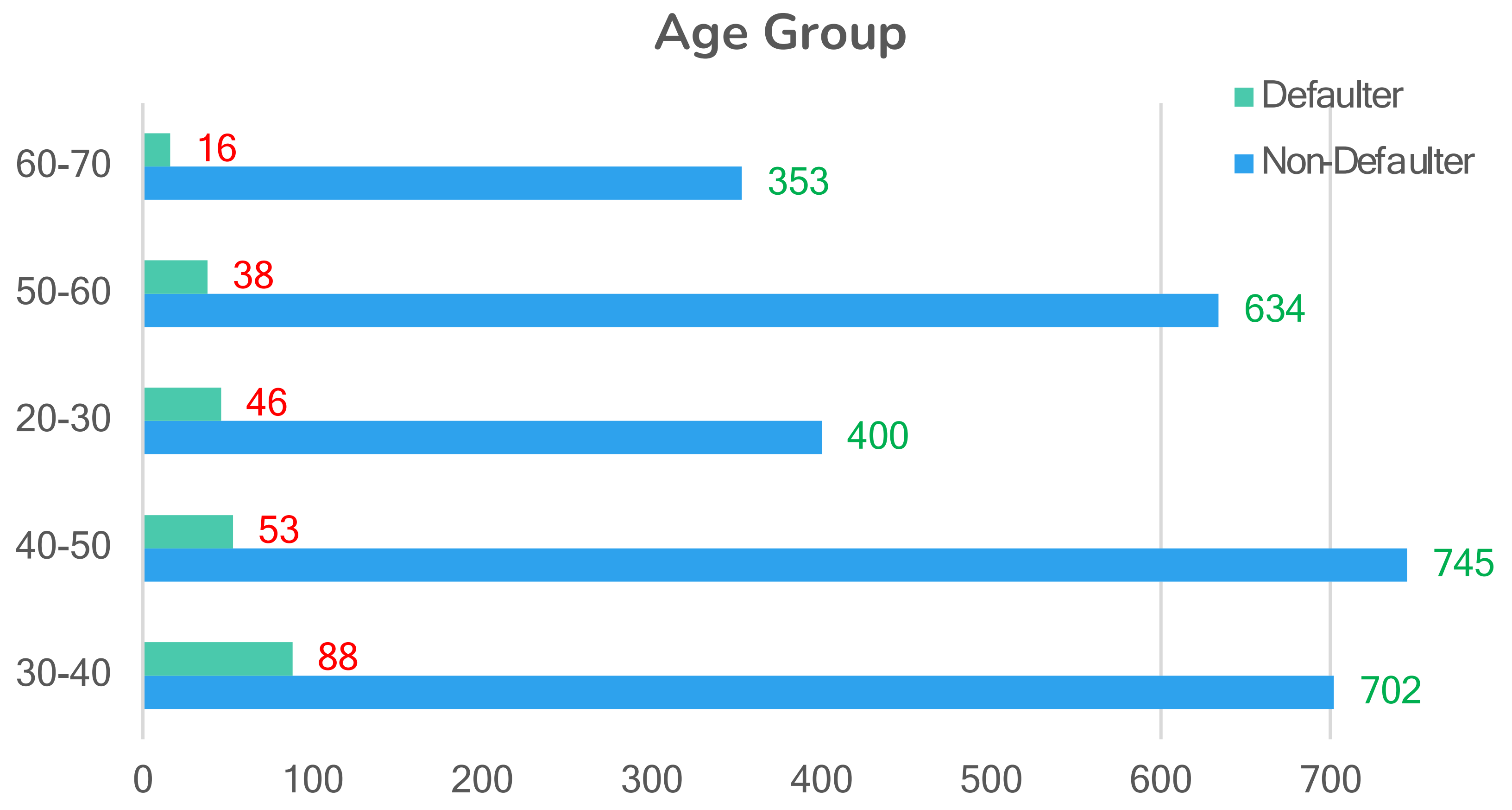
- Client with region rating 3 and Client with low budget loan have higher chances of being defaulter

UNIVARIATE & SEGMENTED UNIVARIATE ANALYSIS



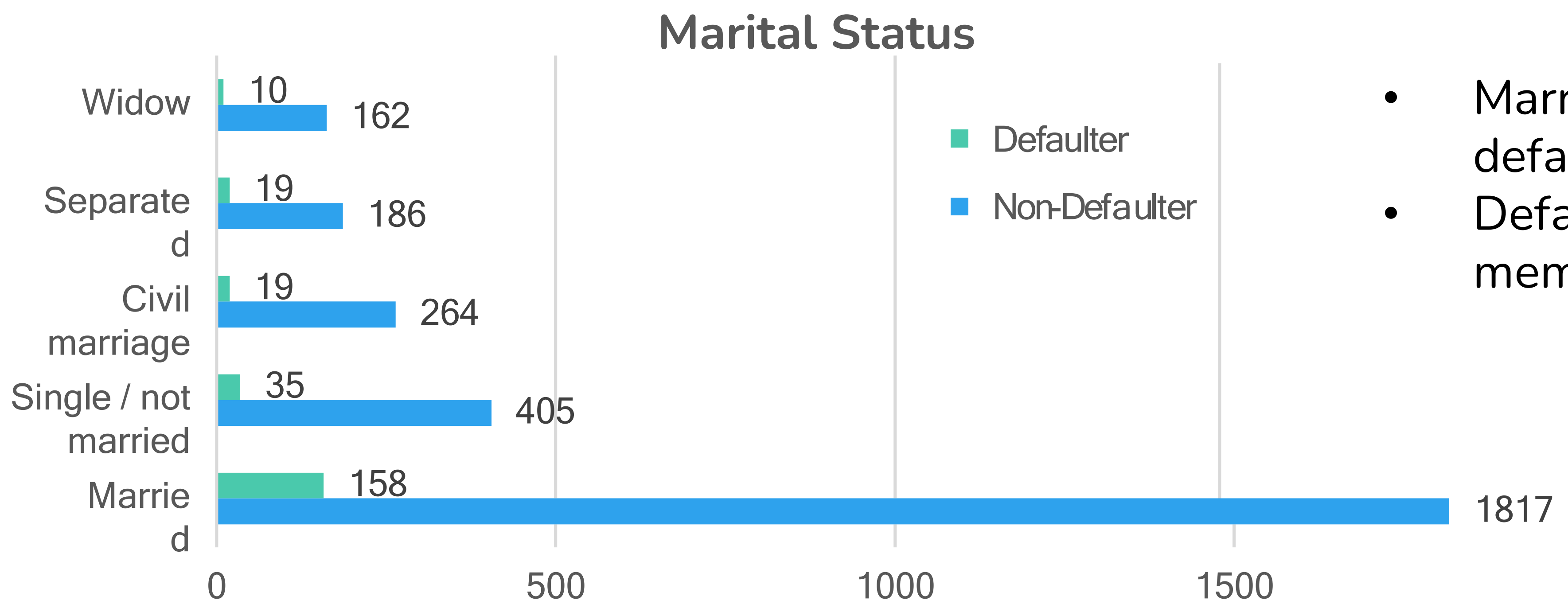
- The less experience the higher chance of defaulter

UNIVARIATE & SEGMENTED UNIVARIATE ANALYSIS

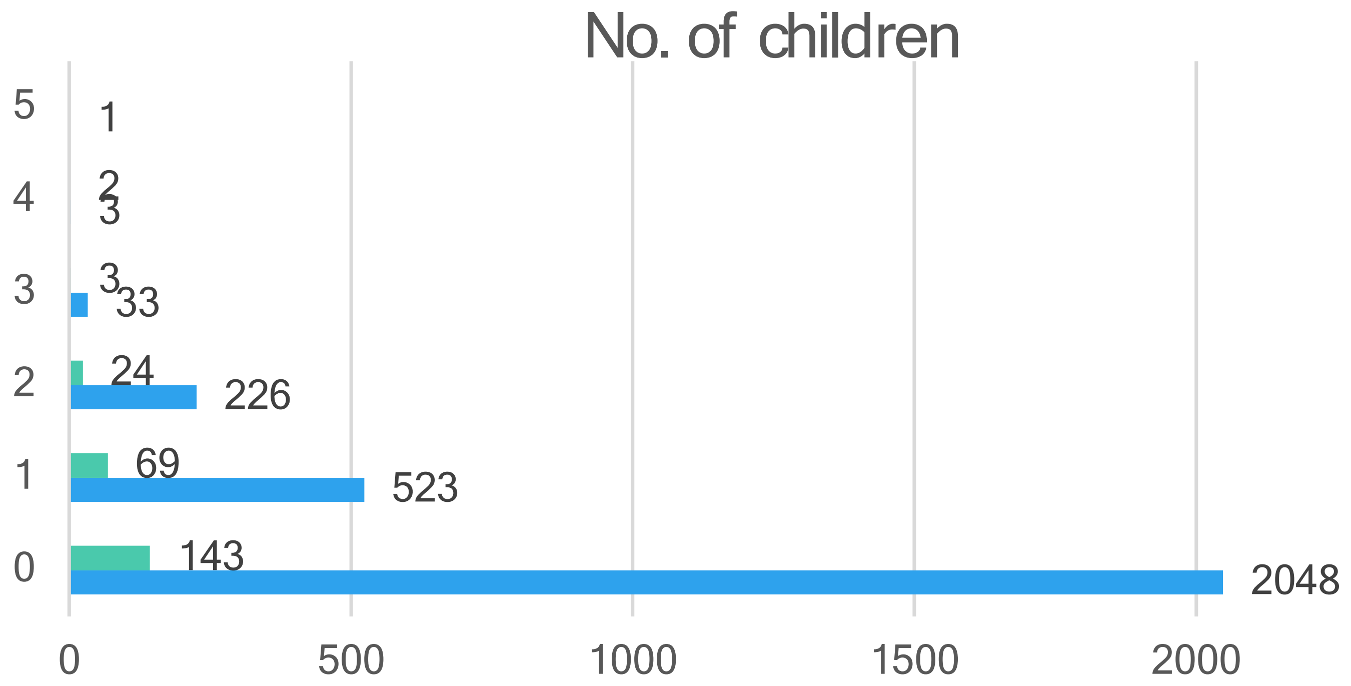


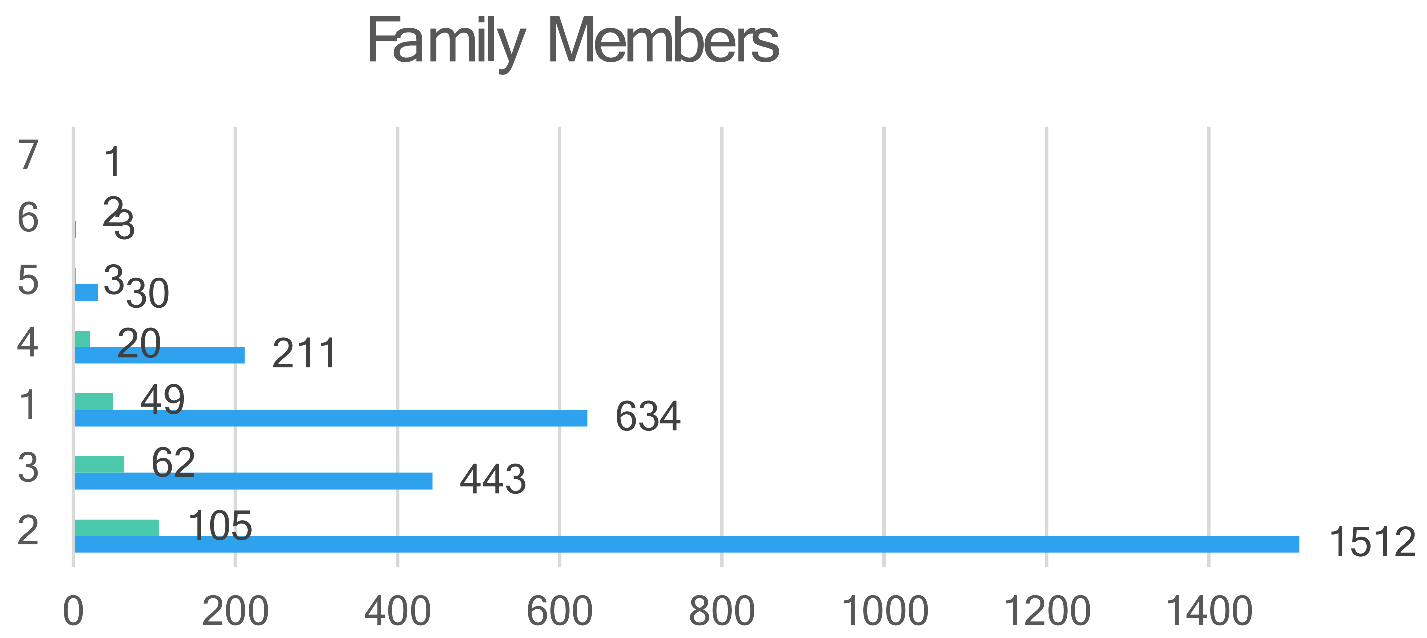
- Client within Age group 20 –40 have max number of defaulter

UNIVARIATE & SEGMENTED UNIVARIATE ANALYSIS

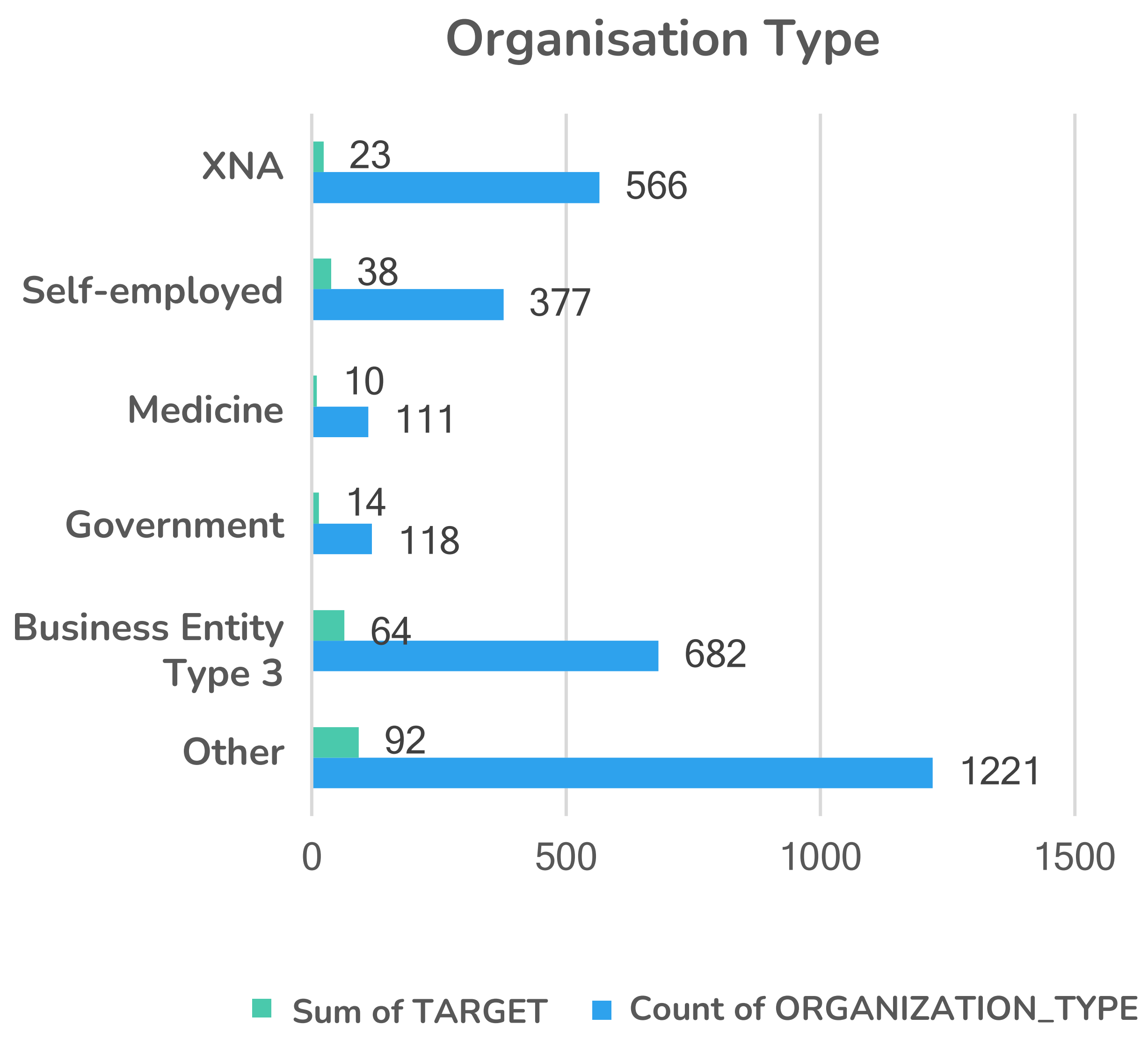


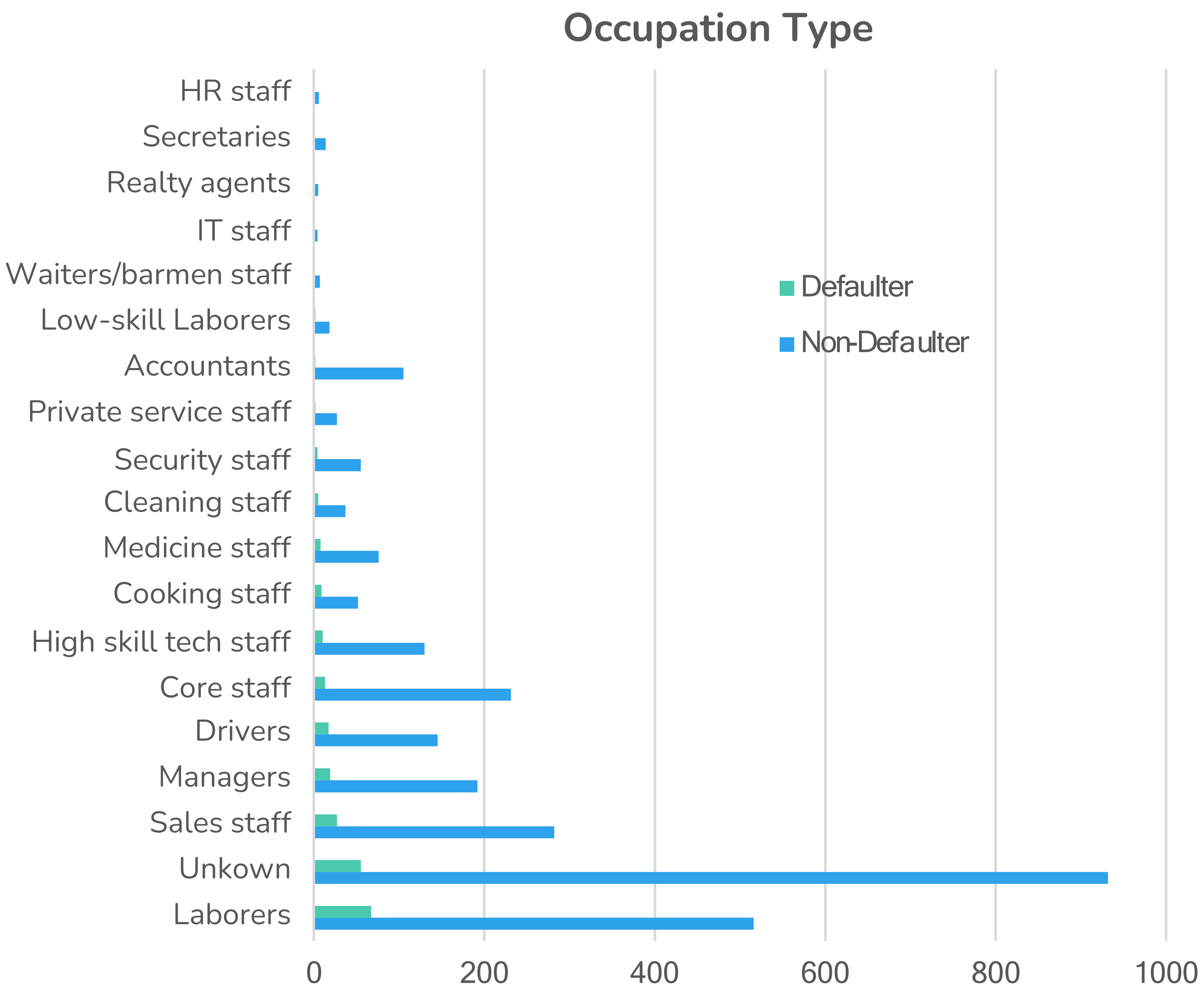
- Married people with 0-1 child have low chances of defaulter
- Defaulter rate increases with increase of family member or no. of children





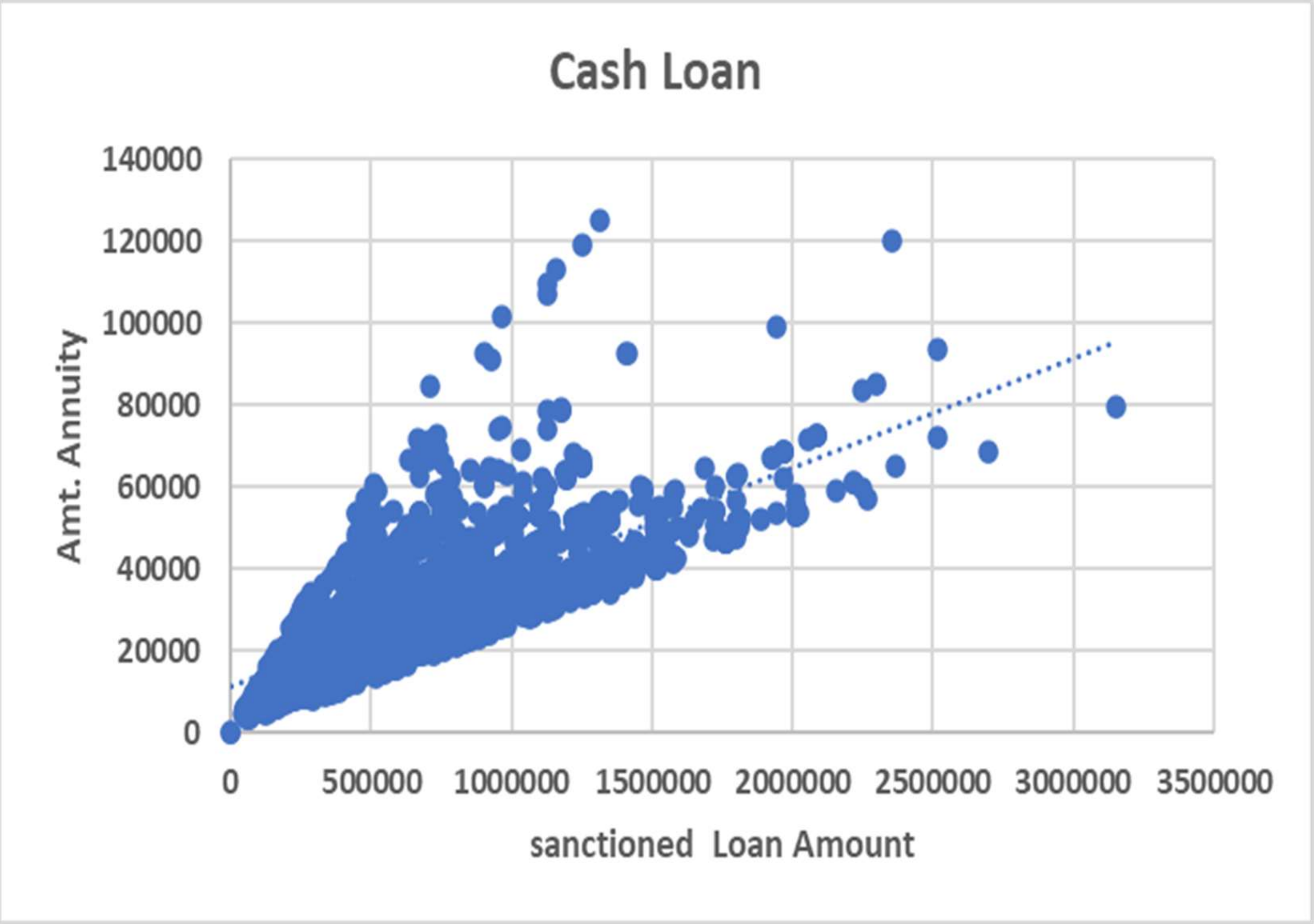
UNIVARIATE & SEGMENTED UNIVARIATE ANALYSIS

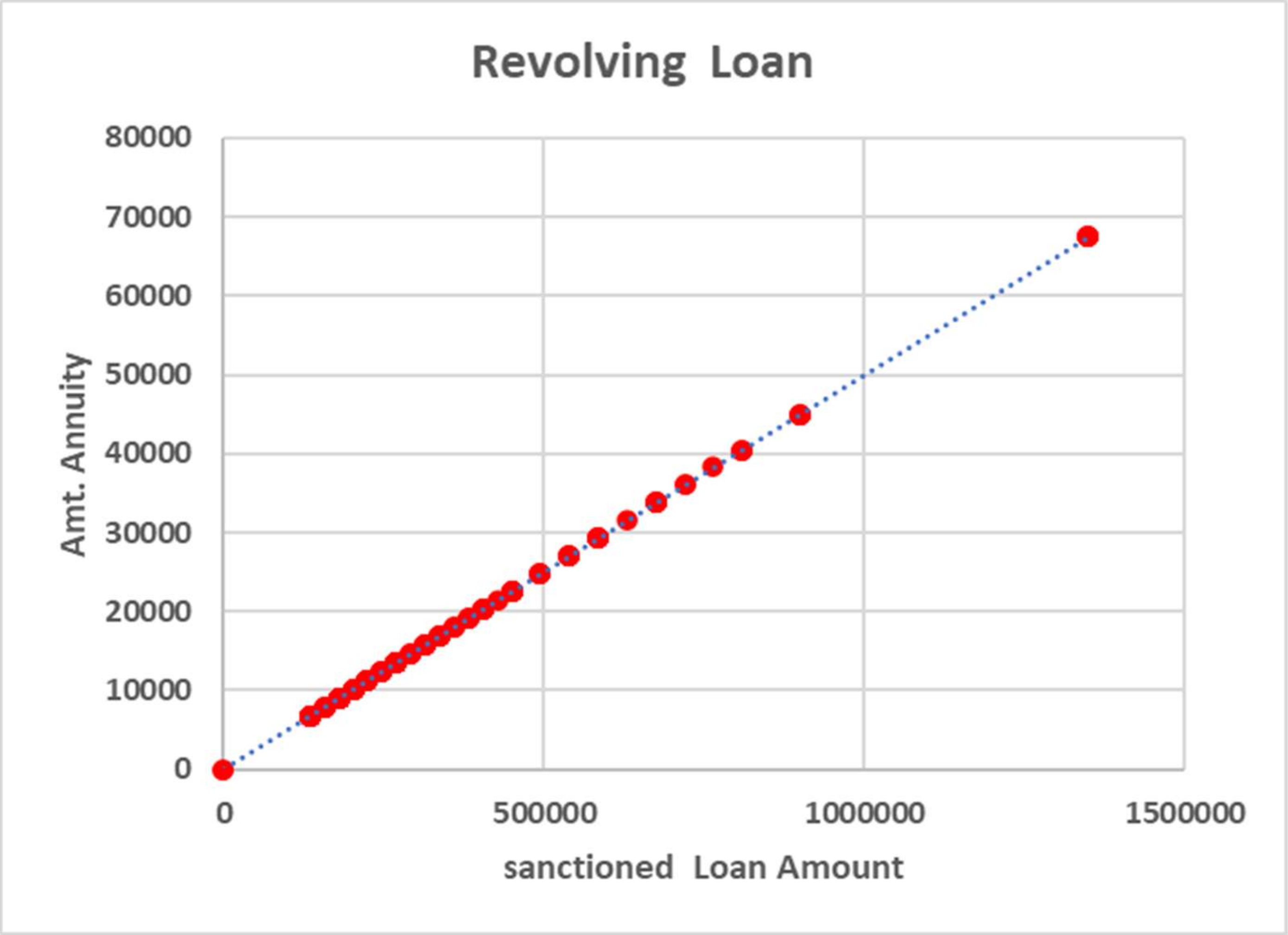




- **Low skilled worker and business Entity type -3 are most risky client**

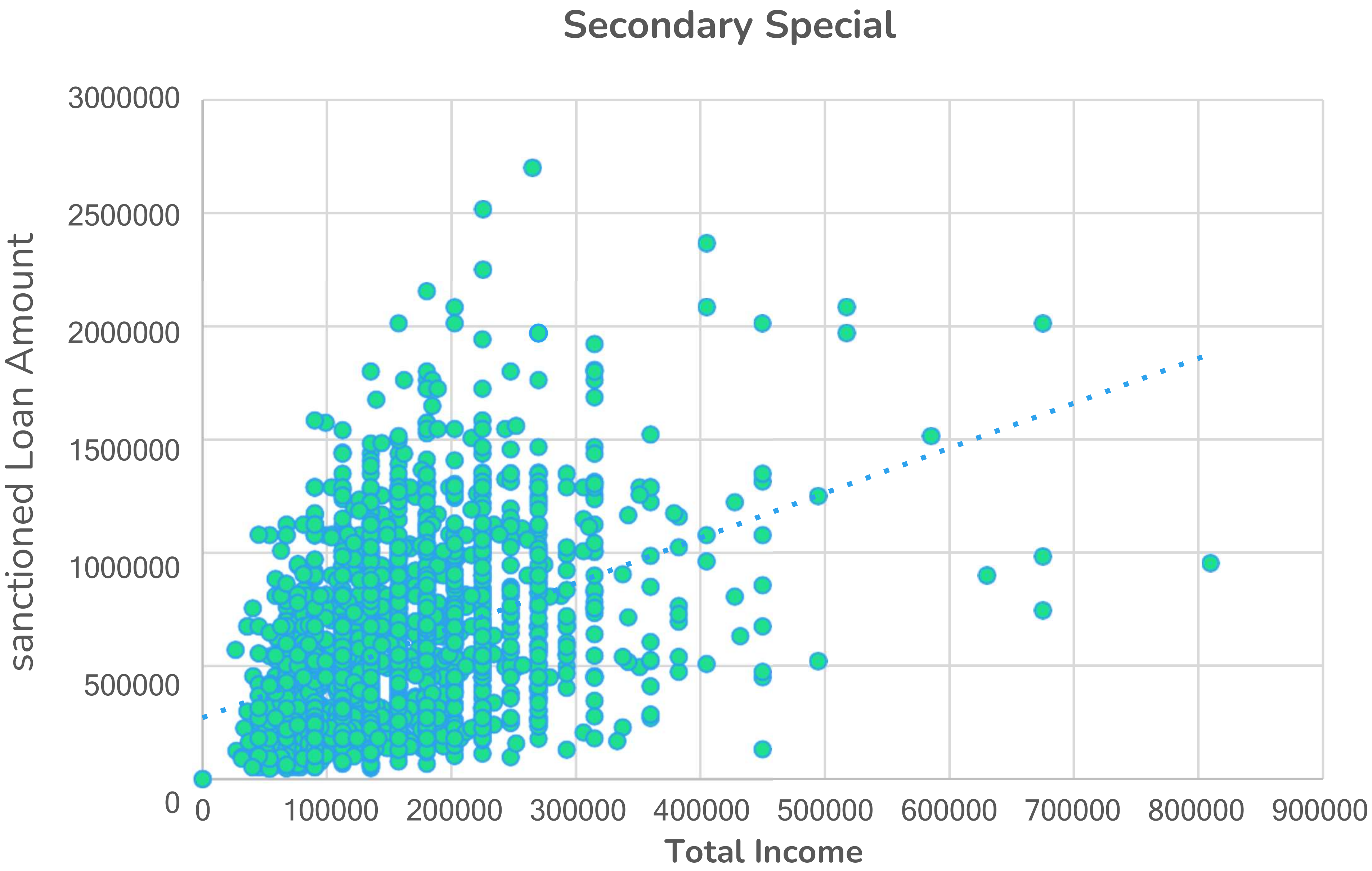
BIVARIATE ANALYSIS

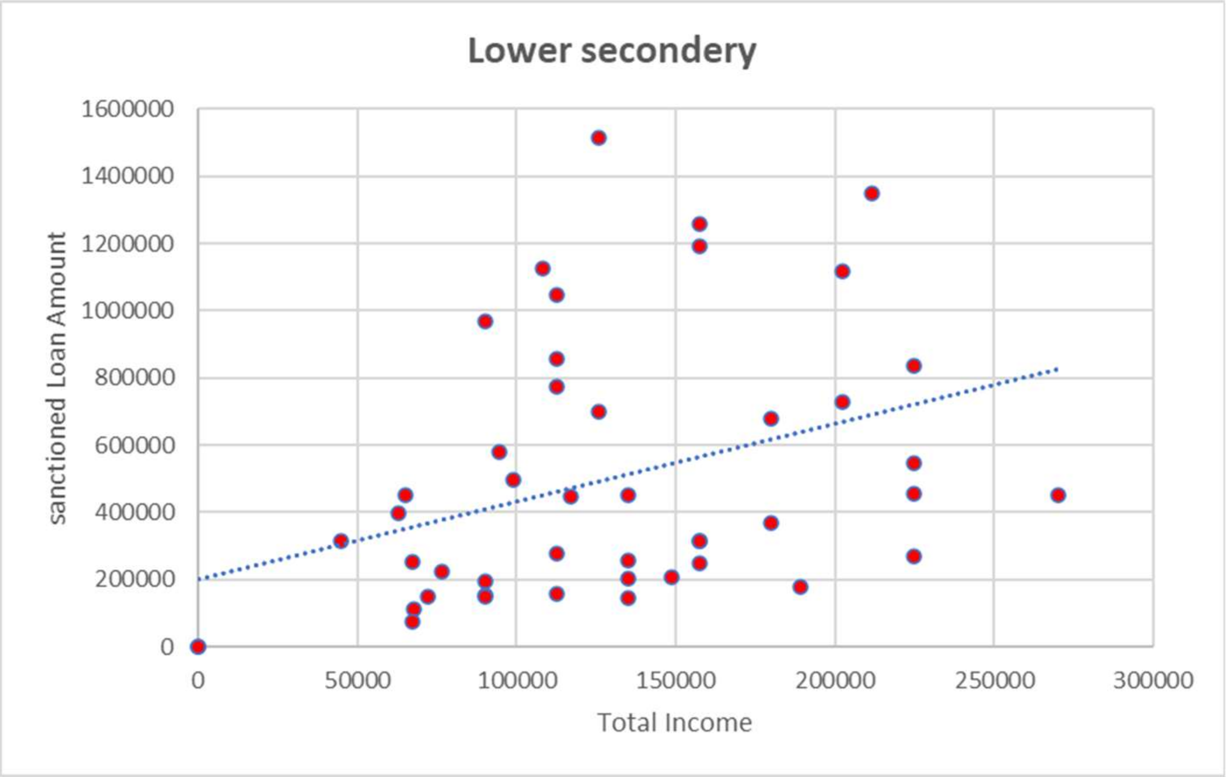




- Loan amount and EMI is highly corelated to each other
- Cash loan also have positive liner graph

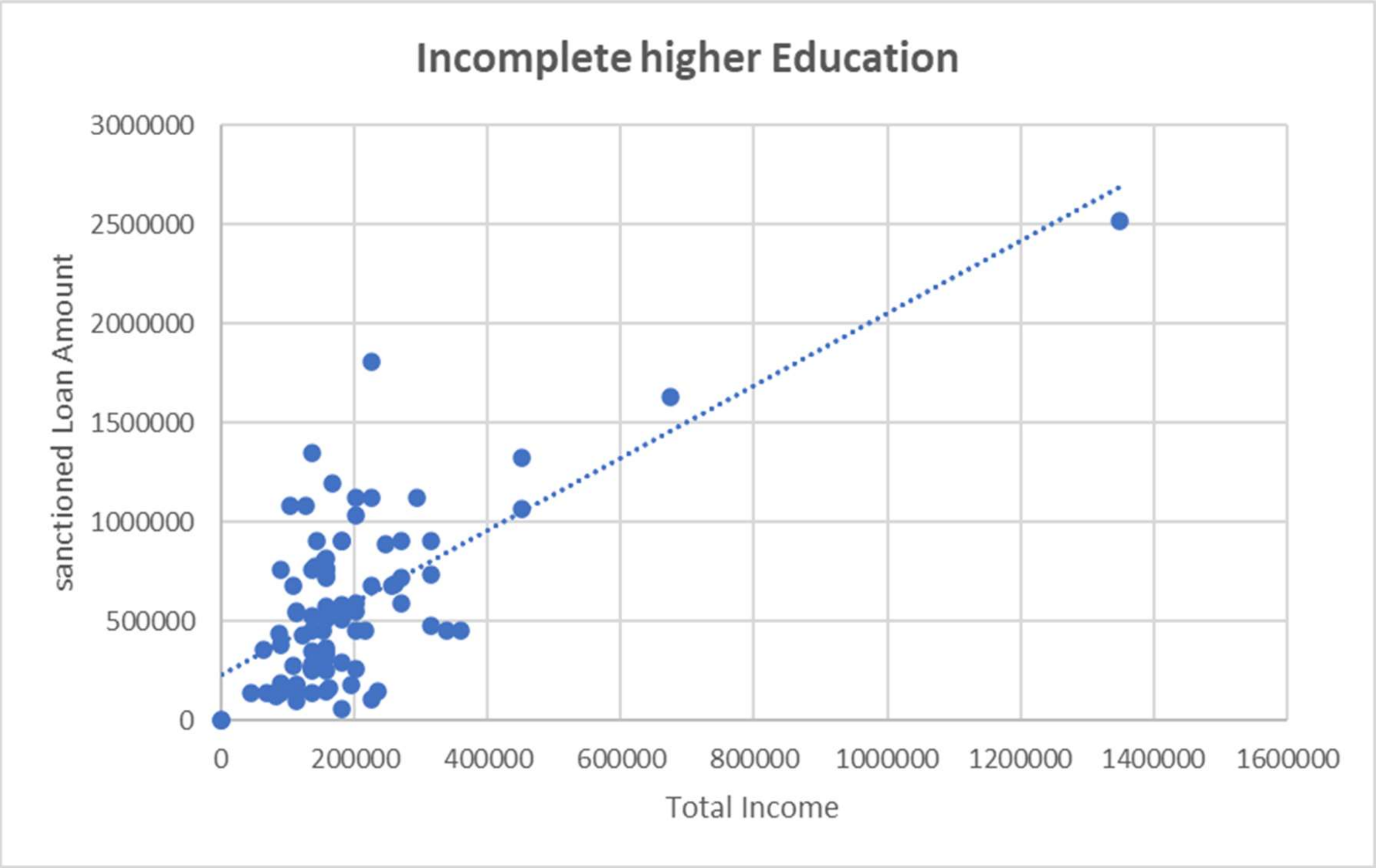
BIVARIATE ANALYSIS

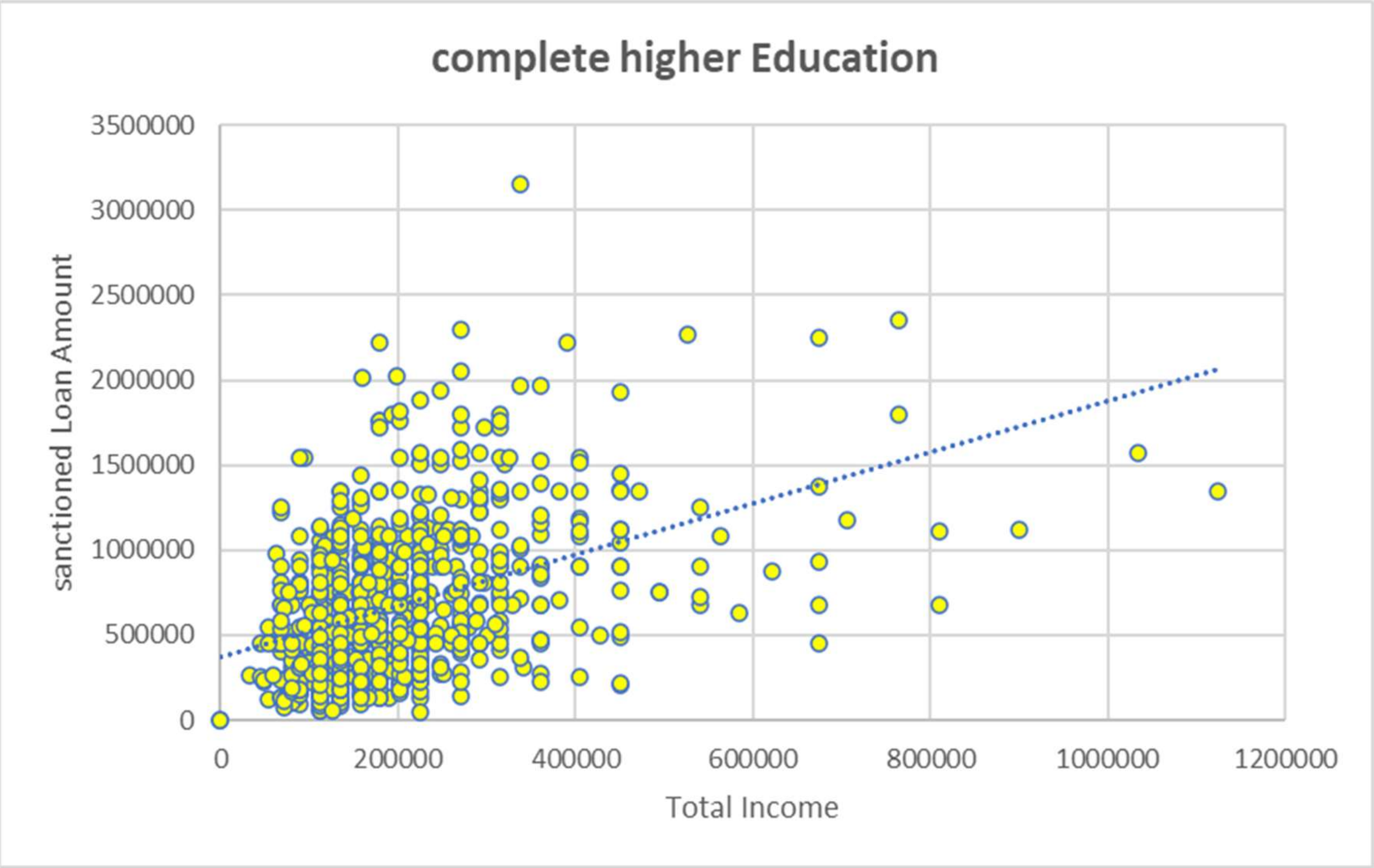




- Client with education secondary and complete higher have many similarities in graph
- While lower secondary class have spread data points

BIVARIATE ANALYSIS





- Excluding outliers in Incomplete higher education class most clients apply for low budget loans

TOP CORRELATED VARIABLE

Top Correlation variable in defaulter vs Non Defaulter

Rank	Var 1 vs Var 2	deflter corr	non-defltr corr
1	AMT_REQ_CREDIT_BUREAU_WEEK vs AMT_REQ_CREDIT_BUREAU_QRT	1. 00	0.00
2	OBS_30_CNT_SOCIAL_CIRCLE vs DEF_30_CNT_SOCIAL_CIRCLE	0.99	0.29
3	AMT_GOODS_PRICE vs AMT_CREDIT	0.98	0.99
4	CNT_FAM_MEMBERS vs CNT_CHILDREN	0.90	0.87
5	DEF_60_CNT_SOCIAL_CIRCLE vs months_LAST_PHONE_CHANGE	0.85	0.00
6	AMT_REQ_CREDIT_BUREAU_MON vs AMT_REQ_CREDIT_BUREAU_YEAR	0.81	0.01
7	AMT_CREDIT vs AMT_ANNUITY	0.77	0.77
8	AMT_ANNUITY vs AMT_GOODS_PRICE	0.77	0.77
9	AMT_REQ_CREDIT_BUREAU_HOUR vs AMT_REQ_CREDIT_BUREAU_DAY	0.74	0.16
10	AMT_CREDIT vs AMT_INCOME_TOTAL	0.51	0.42

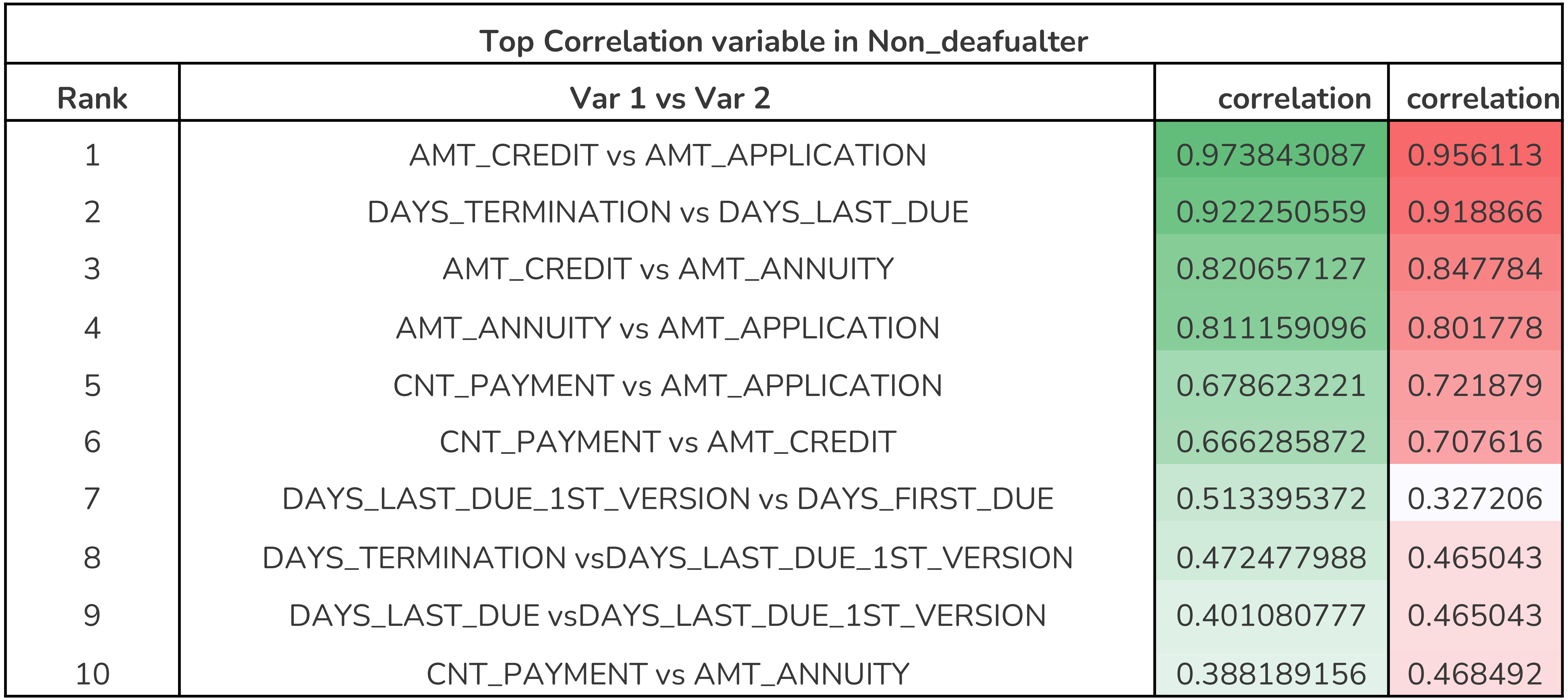
TOP CORRELATED VARIABLE

Top 10 correlated variable in application dataset

Top Correlation variable in Nondefaulter vs defaulter			
Rank	Var 1 vs Var 2	non-defltr corr	defltr corr
1	OBS_60_CNT_SOCIAL_CIRCLE vs OBS 30 CNT SOCIAL CIRCLE	1.00	-0.03
2	AMT_GOODS_PRICE vs AMT_CREDIT	0.99	0.98
3	CNT_FAM_MEMBERS vs CNT_CHILDREN	0.87	0.90
4	DEF_60_CNT_SOCIAL_CIRCLE vs DEF_30_CNT_SOCIAL_CIRCLE	0.86	-0.15
5	AMT_CREDIT vs AMT_ANNUITY	0.77	0.77
6	AMT_ANNUITY vs AMT_GOODS_PRICE	0.77	0.77
7	AGE vs Exp. In Years	0.62	0.55
8	AMT_ANNUITY vs AMT_INCOME_TOTAL	0.51	0.49
9	AMT_GOODS_PRICE vs AMT_INCOME_TOTAL	0.43	0.50
10	AMT_CREDIT vs AMT_INCOME_TOTAL	0.42	0.51

TOP CORRELATED VARIABLE

Top 10 correlated variable in Previous application dataset



Top Correlation variable in Non_deafualter			
Rank	Var 1 vs Var 2	correlation	correlation
1	AMT_CREDIT vs AMT_APPLICATION	0.973843087	0.956113
2	DAYS_TERMINATION vs DAYS_LAST_DUE	0.922250559	0.918866
3	AMT_CREDIT vs AMT_ANNUITY	0.820657127	0.847784
4	AMT_ANNUITY vs AMT_APPLICATION	0.811159096	0.801778
5	CNT_PAYMENT vs AMT_APPLICATION	0.678623221	0.721879
6	CNT_PAYMENT vs AMT_CREDIT	0.666285872	0.707616
7	DAYS_LAST_DUE_1ST_VERSION vs DAYS_FIRST_DUE	0.513395372	0.327206
8	DAYS_TERMINATION vsDAYS_LAST_DUE_1ST_VERSION	0.472477988	0.465043
9	DAYS_LAST_DUE vsDAYS_LAST_DUE_1ST_VERSION	0.401080777	0.465043
10	CNT_PAYMENT vs AMT_ANNUITY	0.388189156	0.468492

- **In previous dataset we can not see any major difference while it compare to Application dataset**

INTERFERENCES

Decisive Factor whether an applicant will be Defaulter:

- CODE_GENDER: Men are at relatively higher default rate
- NAME_FAMILY_STATUS : Married people with 0-1 child have low chances of defaulter rate increases with increase of family member or no. of children
- Defaulter rate increases with increase of family member or no. of children
- NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education have more defaulter rate
- REGION_RATING_CLIENT: People who live in **Rating 3** has highest defaults.
- OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as the default rate is huge.
- ORGANIZATION_TYPE: **Business Entity type – 3** have highest rate of default
- DAYS_BIRTH: Avoid young people who are in **age group of 20-40** as they have higher probability of defaulting
- DAYS_EMPLOYED: The less experience the higher chance of defaulter
- CNT_CHILDREN & CNT_FAM_MEMBERS: Client who have higher number of children increases default rate.

Note – All this inferences is taken from analysis of sample dataset it will be not applicable to original dataset

SUMMARY

- Gender , Income Type, Region , education and working experience factors are most important while approving loan
- Client with low income Category and Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff have higher defaulter chances so here Bank have opportunity to approve loan with higher rate of Interest
- Educated client less likely to being defaulter
- Client with working experience under 10 years and within age group 20 to 40 are risky so bank should lend them loan on higher risk
- Client who changes mobile number within 2 months have higher defaulter rate

CONCLUSION

- This Case Study is aims to application of exploratory data analysis techniques on dataset to finding meaningful insights.
- It build concept of use of Univariate Analysis and Bi/Multivariate Analysis.
- This case study give an idea of risk Analytics

THANK YOU