Credit Bank-Loan: Exploratory Data Analysis

-Vivek Kumar

Introduction to the Problem

When the company receives a loan application, the company has to decide for loan approval based on the applicants profile. Two types of risks are associated with the banks 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.

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

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

Objective of the Analysis

This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.
To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

Strategic analysis techniques employed:

Strategic analysis

This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

01

MANOVA technique

compared several groups with respect to multiple continuous variables

02

Univariate analysis

it explores each variable separately

03

Bivariate analysis

the analysis of two variables, for the purpose of determining the empirical relationship between them.

04

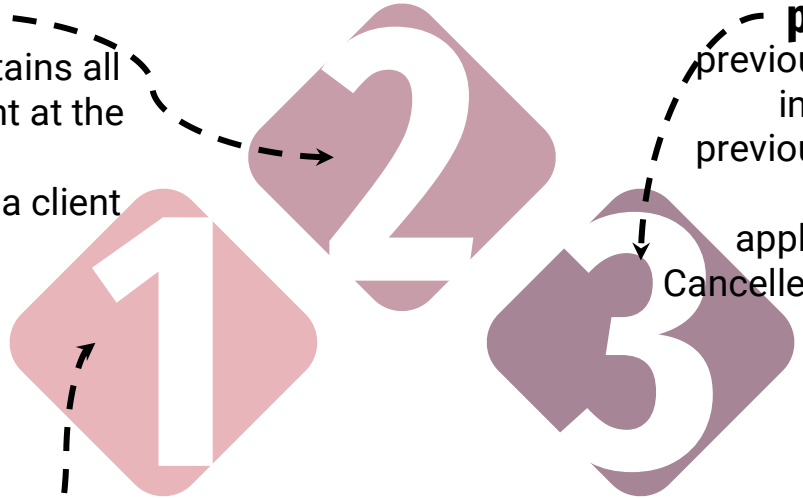
Multivariate analysis

the simultaneous observation and analysis of more than one outcome variable.

Datasets given for analysis

Application_data.csv

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

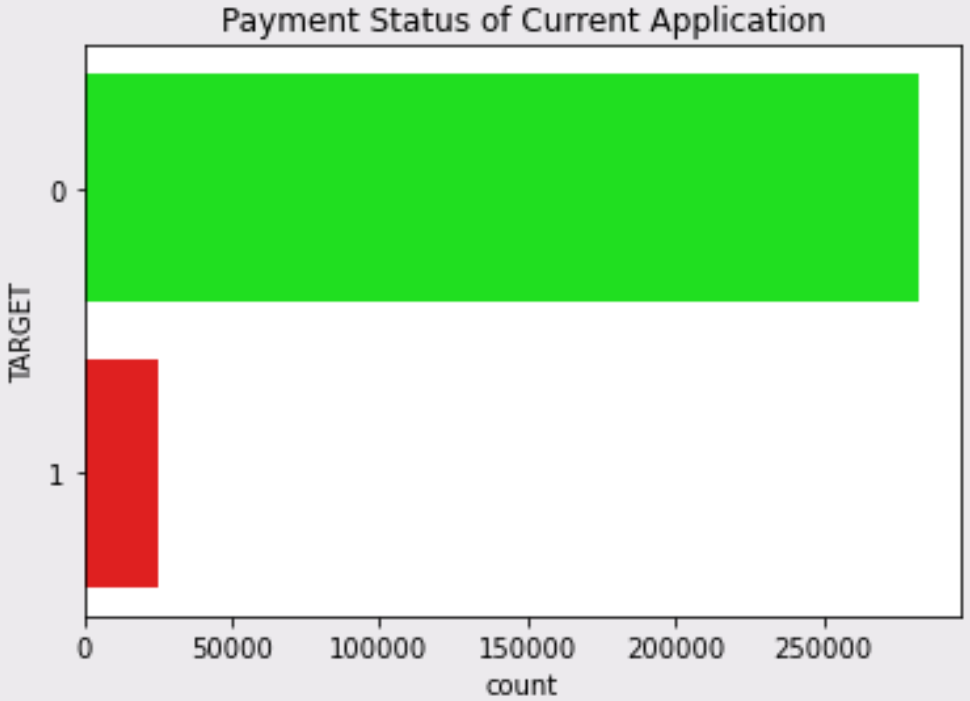
previous_application.csv contains information about the clients previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

columns_description.csv

columns_description.csv is data dictionary which describes the meaning of the variables.

Identifying if the data is misproportioned:

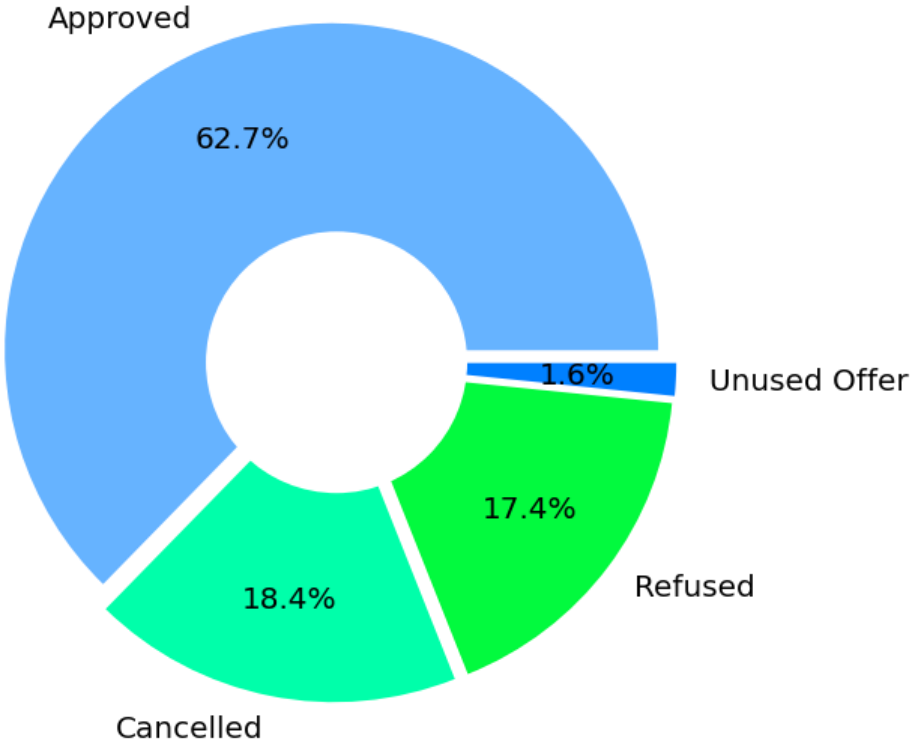
*We can identify that by examining the repayment trend as per **TARGET** Colum*



Examining the repayment trend as per TARGET column:

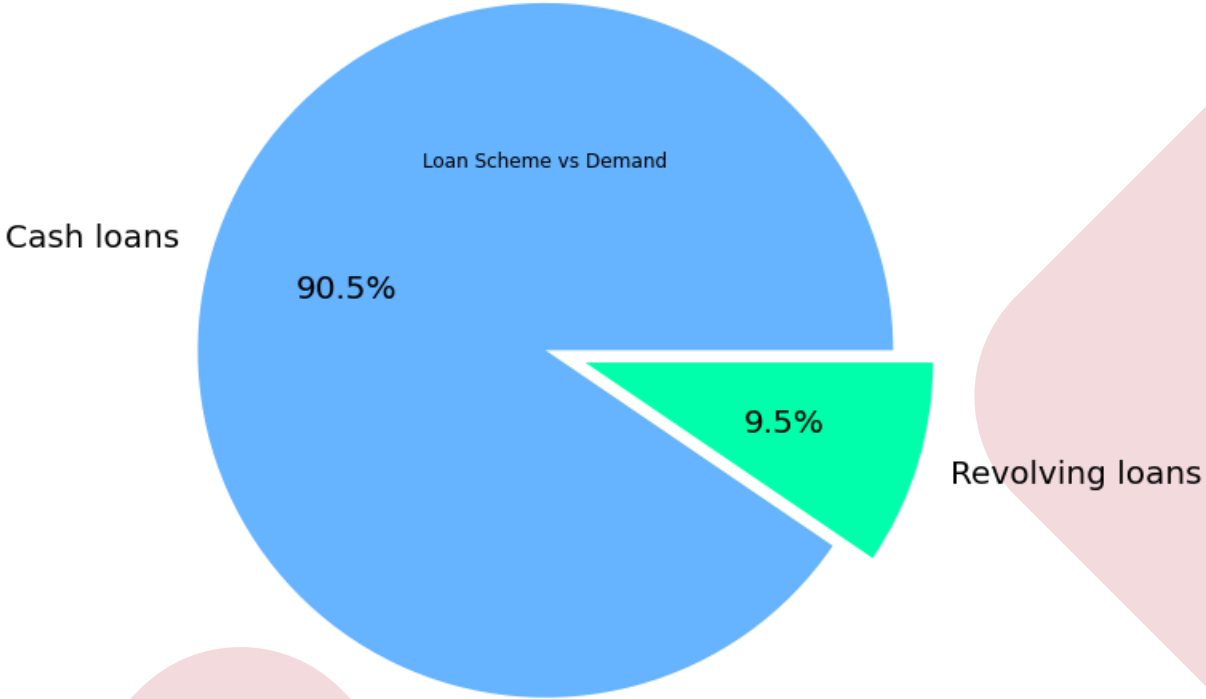
***NOTE:**TARGET column is 1 for defaulters & 0 otherwise*

Payment Status of Previous Application & Payment



Examining the repayment trend as per NAME_CONTRACT_STATUS column

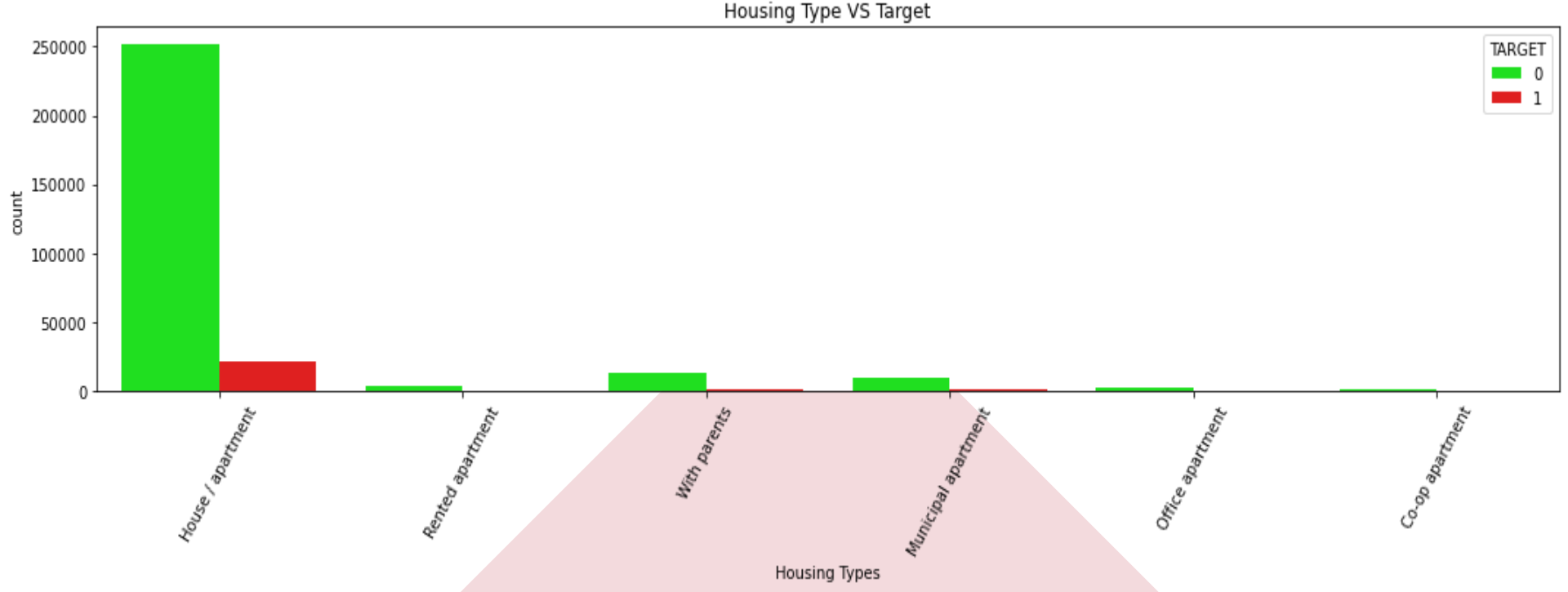
Most common type of loan that customers are intrested in:



We can get it by looking at the countplot of ***NAME_CONTRACT_TYPE***

Analysis based on customers residence

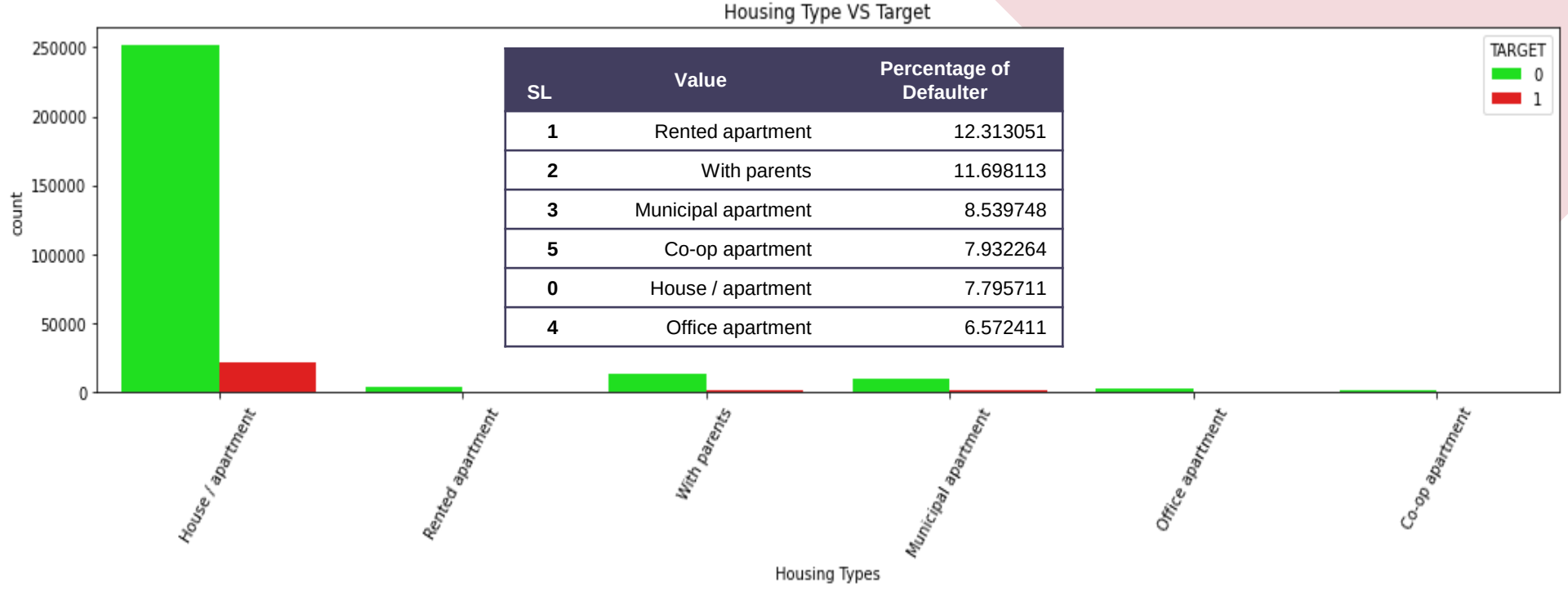
- *We can easily observe that **repayment rate** increases with increase in **defaulter rate**.*
- *Most of the customers live in **House/Apartment***
- *Most of the defaulters also live in **House/Apartment***
- *Rented apartments have higher rate of defaulters as they can always shift(as per the data)*



Analysis based on customers residence

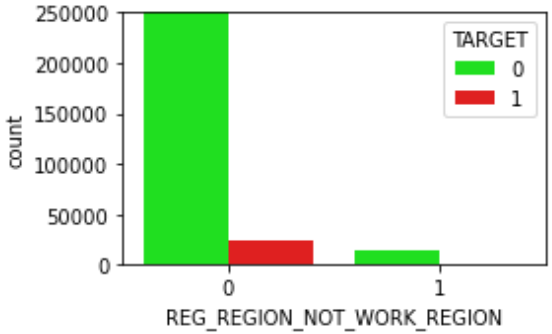
- *We can easily observe that **repayment rate** increases with increase in **defaulter rate**.*
- *Most of the customers live in **House/Apartment***
- *Most of the defaulters also live in **House/Apartment***
- *Rented apartments have higher rate of defaulters as they can always shift(as per the data)*

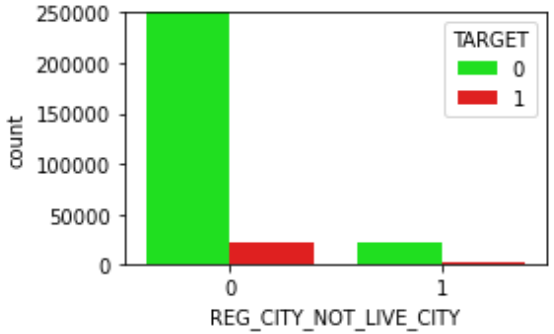
SL	Value	Percentage of Defaulter
1	Rented apartment	12.313051
2	With parents	11.698113
3	Municipal apartment	8.539748
5	Co-op apartment	7.932264
0	House / apartment	7.795711
4	Office apartment	6.572411

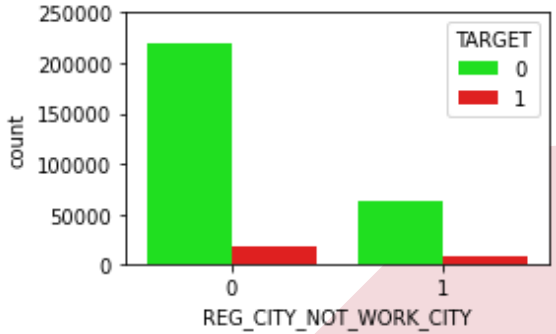


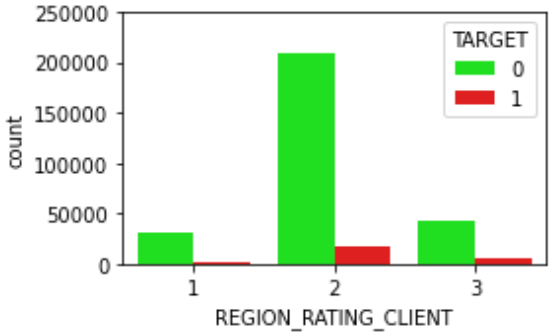
Analysis based on customers region

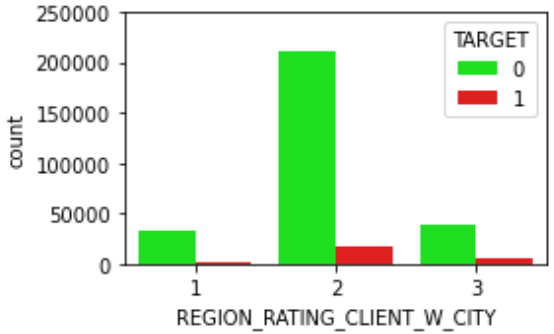
- *We can easily observe that **defaulter rate** increases when:*
  1. *Current region is same as working region.*
  2. *Current city is same as city in which client is living in.*
  3. *Current city is same as city in which client is working in.*
- *Most of the customers live in Region rated **2nd** among 3 regions.*
- *Defaulters are the most in **region 2** then they decrease in **region 3** and least are in **region 1**.*
- *Increase in defaulters per region is natural because the number of customers also increases.*





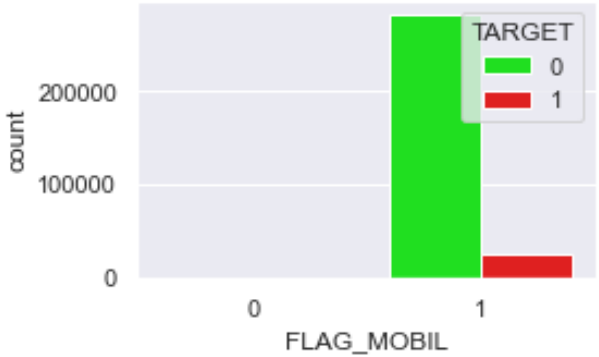


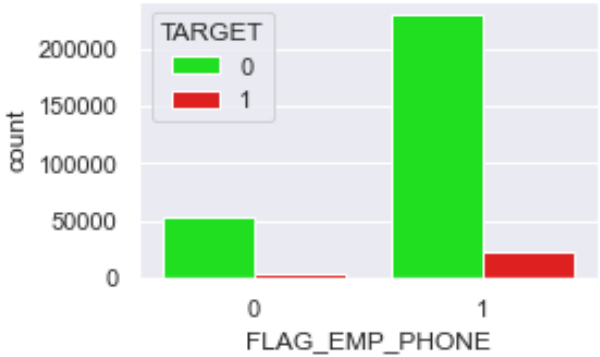


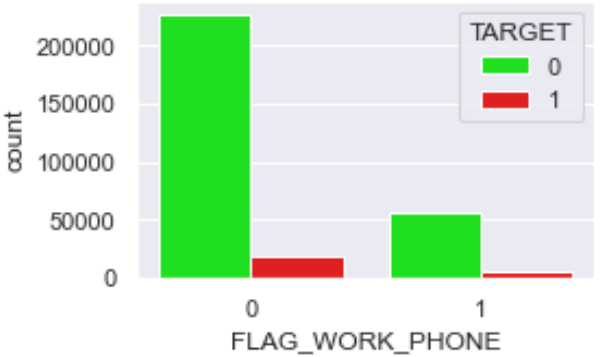


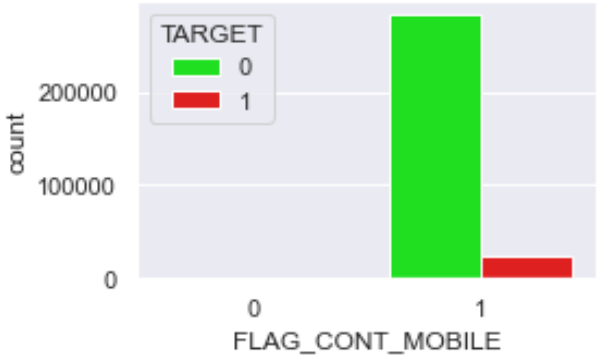
Analysis based on customers contact details

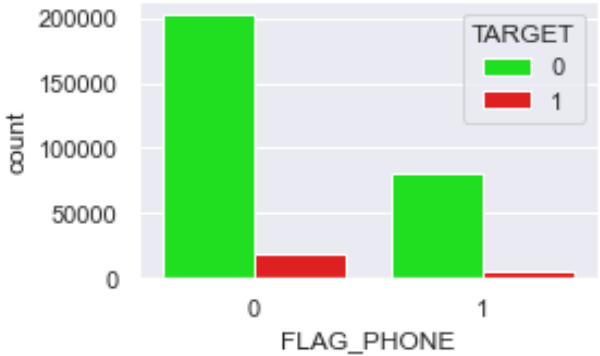
- *We can easily observe that **defaulters** havent provided:*
  1. *Their Email address(es).*
  2. *Their Home phone.*
- ***FLAG_PHONE** and **FLAG_WORK_PHONE** are identical hence either can be dropped.*

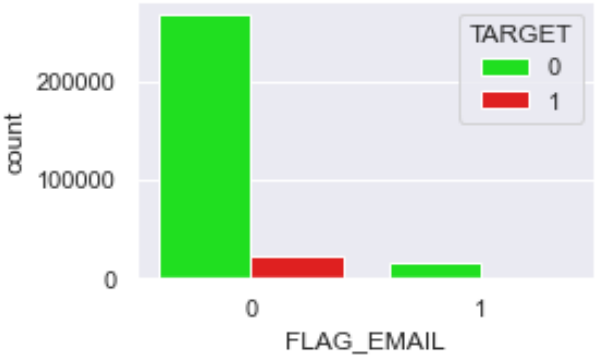






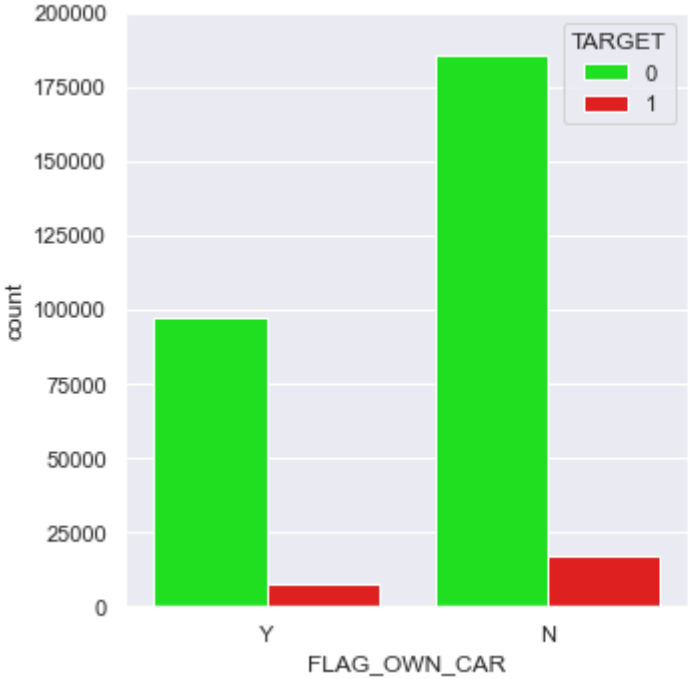


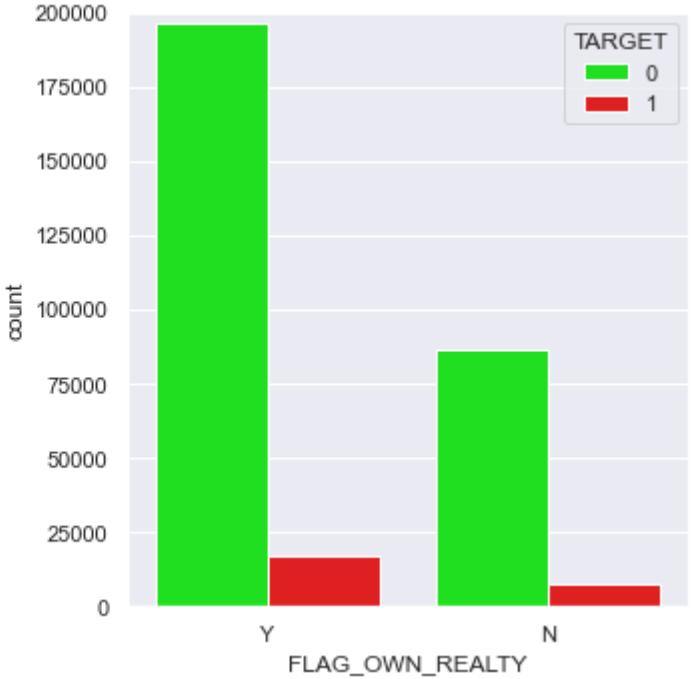




Analysis based on customers asset details/value

- *We can easily observe that defaulters typically dont own a **car**.*
- *Also notice that most of the customers own a **realty** as per the trend.*
- *People not having a **realty** and **car** and have higher chances of default than the people who own **realty** and **car***
- *Defaulter or not, most applicants have car age between 0-25 years.*





FLAG_OWN_CAR

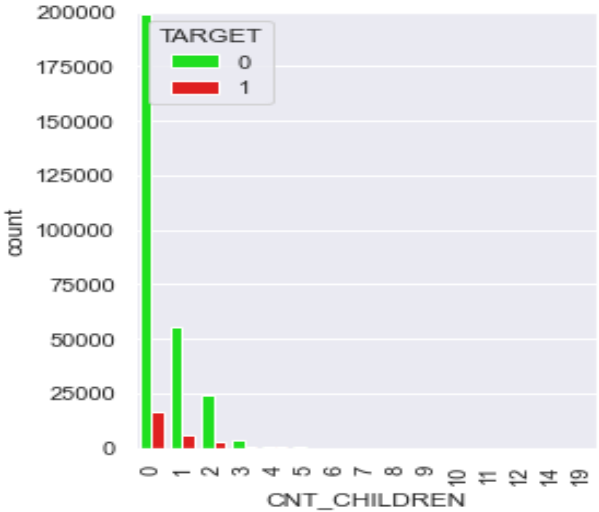
	Value	Percentage of Defaulter
0	N	8.500227
1	Y	7.243730

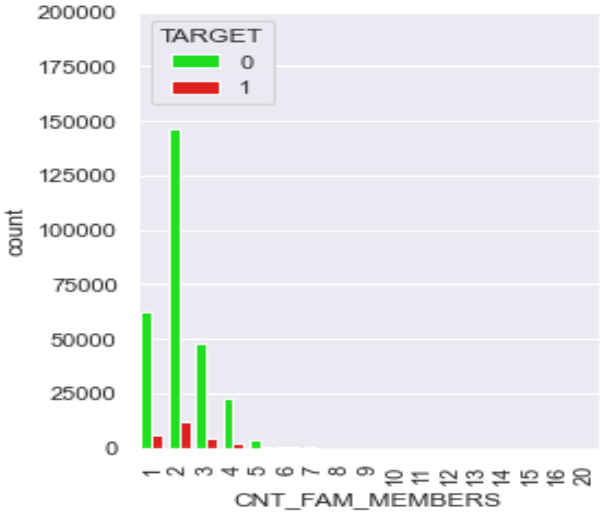
FLAG_OWN_REALTY

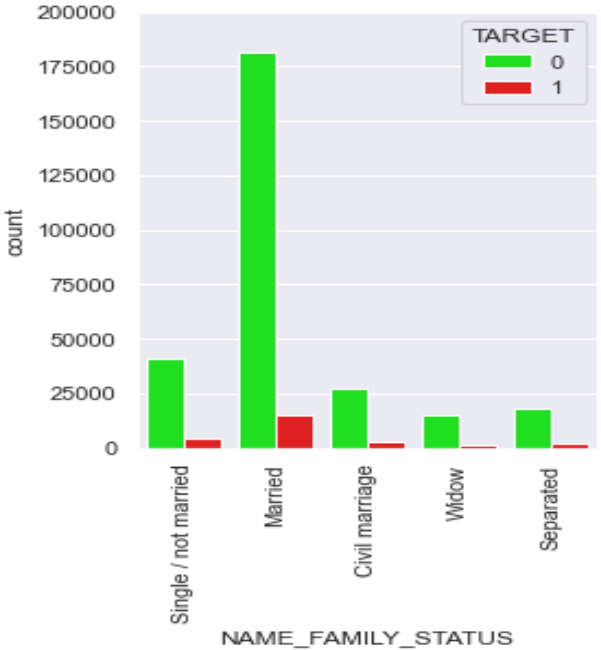
	Value	Percentage of Defaulter
1	N	8.324929
0	Y	7.961577

Analysis on basis of Family Related Info

- *Defaulters are generally under category **Civil Marriage** and **Single applicants**.*
- *Most of the defaulters are having **1-4** family members.*
- *Most of the defaulters are having **0-2** childrens.*
- *Customers having **9** or **11** children can be suspected as outliers in the df.*
- *Customers having **11** or **13** members can be suspected as outliers in the df.*
- *Most of the defaulters were **unacompanied** during Loan screening process.*





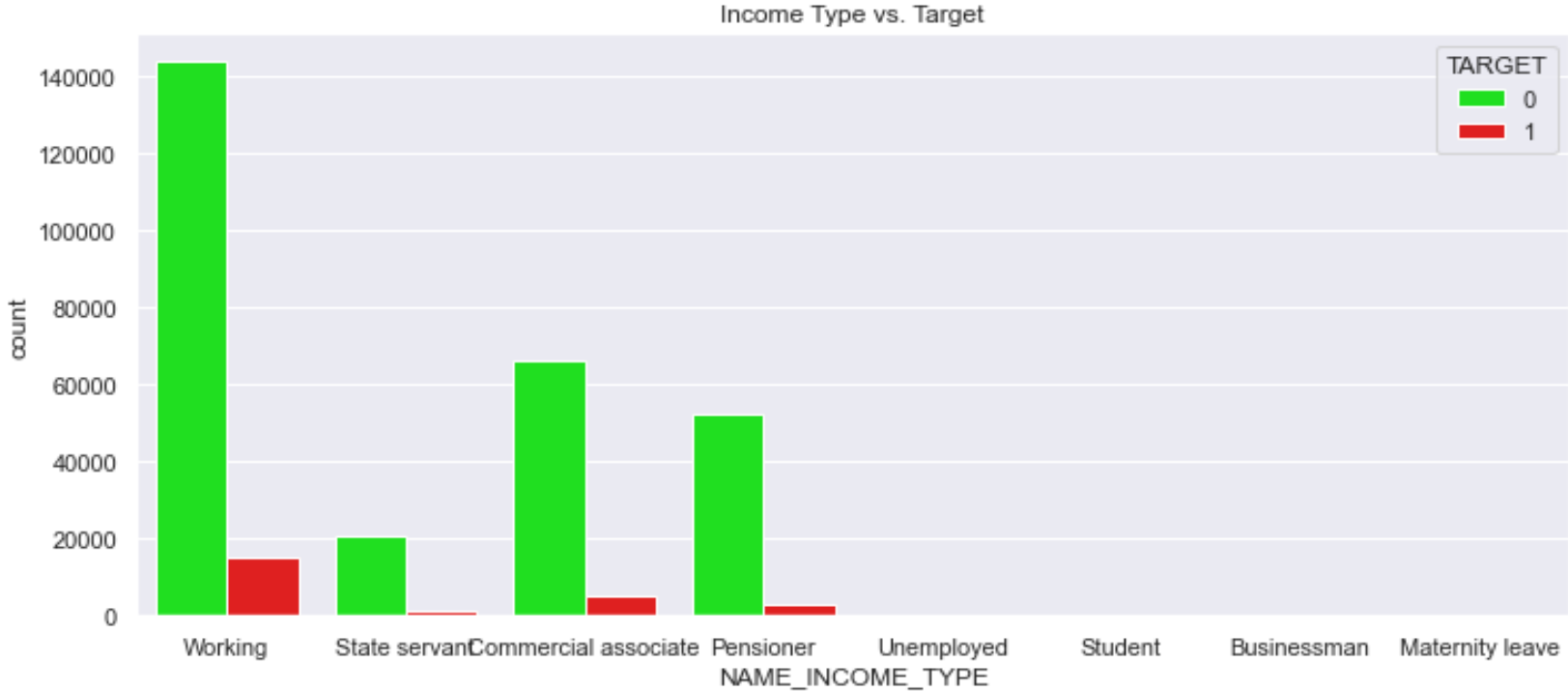


Analysis on basis of Family Related Info (contd.)

- *Defaulters are generally under category **Civil Marriage** and **Single applicants**.*
- *Most of the defaulters are having **1-4** family members.*
- *Most of the defaulters are having **0-2** childrens.*
- *Customers having **9** or **11** children can be suspected as outliers in the df.*
- *Customers having **11** or **13** members can be suspected as outliers in the df.*
- *Most of the defaulters were **unacompanied** during Loan screening process.*

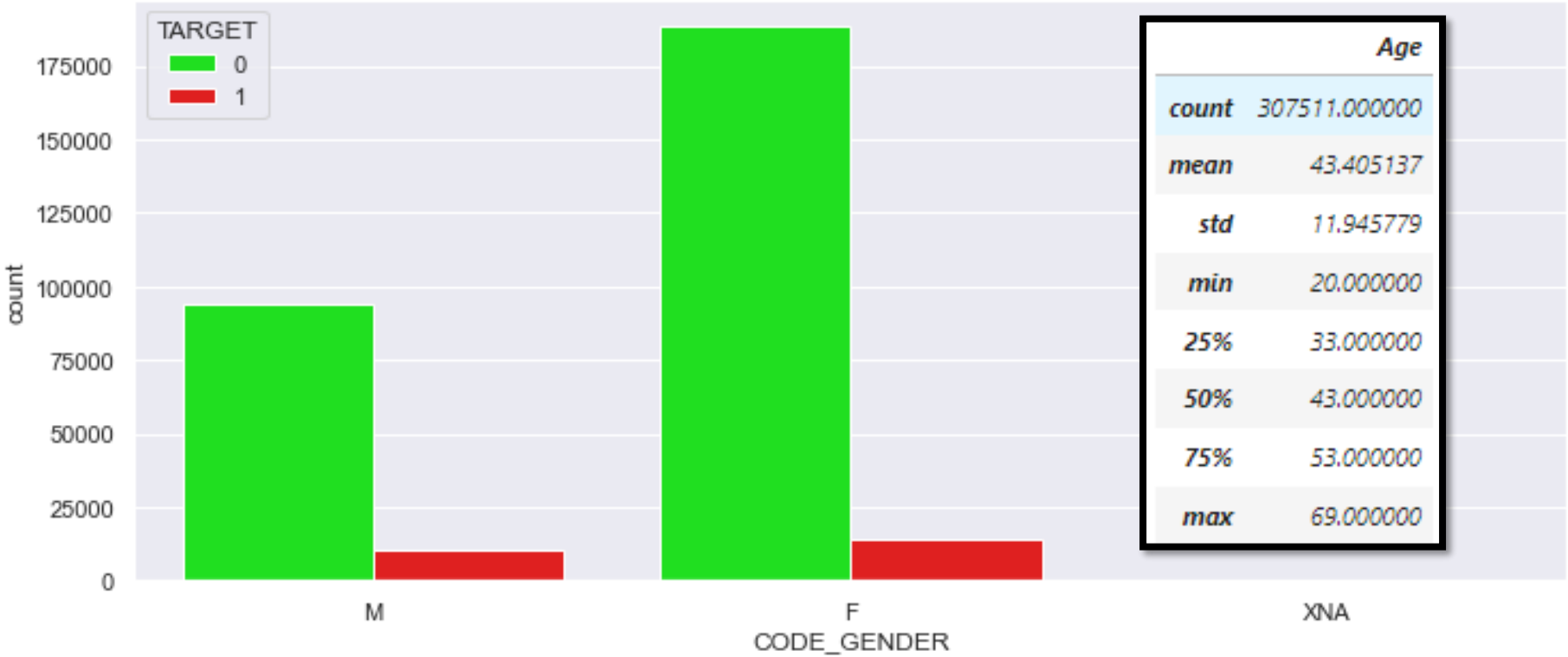
Analysis on basis of Education and Occupation Info

- *Defaulters are generally **Unemployed** or in **Maternity leave**.*
- *Most of the customers are **Working**.*
- *Most of the customers have not completed their **Higher** education.*
- *Most of the workina defaulters are workina as **Low-skilled labourers**.*



Analysis on basis of Customer profile(age, sex etc)

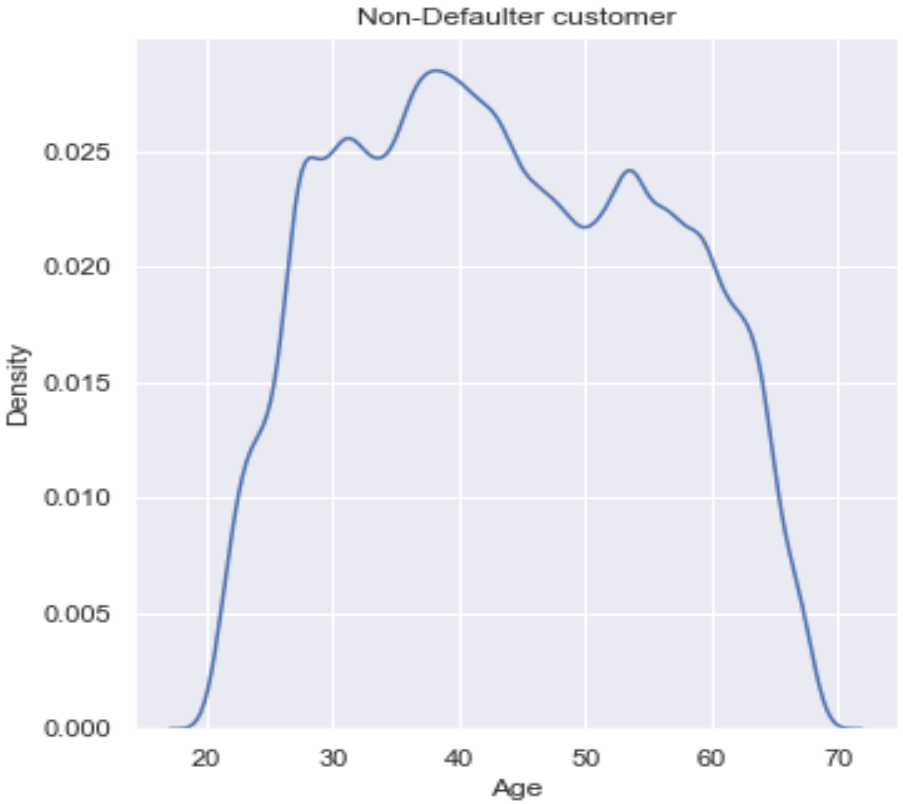
- *The customers are having age between **20** to **69**.*
- *People of age **25-35** have higher default rate.*
- *Default cases are less for applicants more than **40 years** old..*

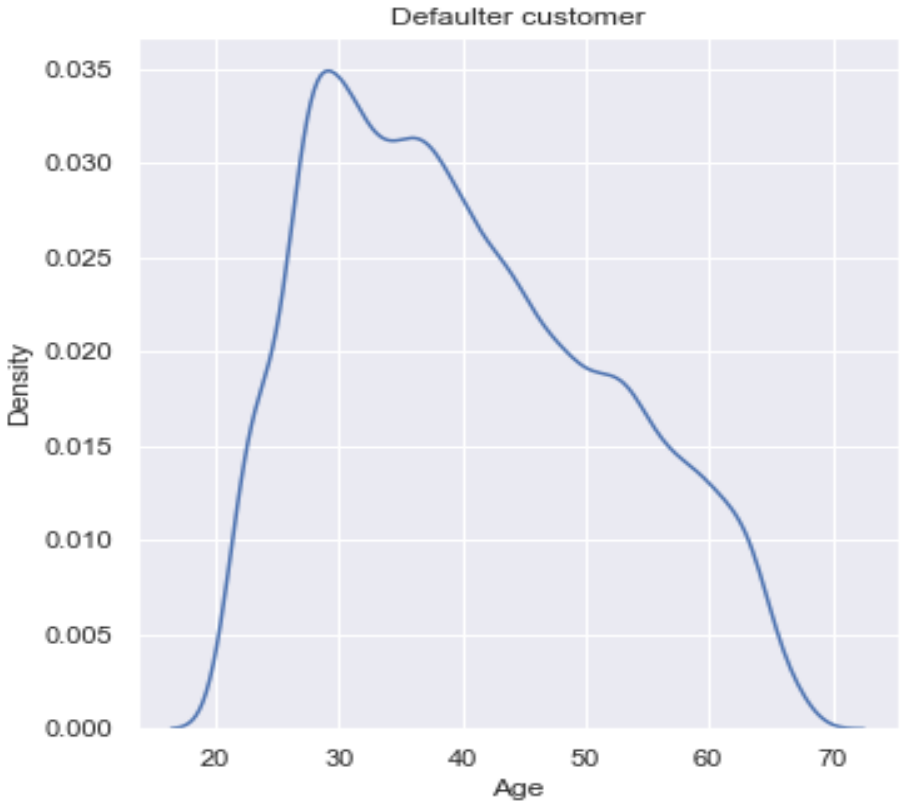


	Age
count	307511.000000
mean	43.405137
std	11.945779
min	20.000000
25%	33.000000
50%	43.000000
75%	53.000000
max	69.000000

Analysis on basis of age

- *People of age **25-35** have higher default rate.*
- *Default cases are less for applicants more than **40 years** old..*





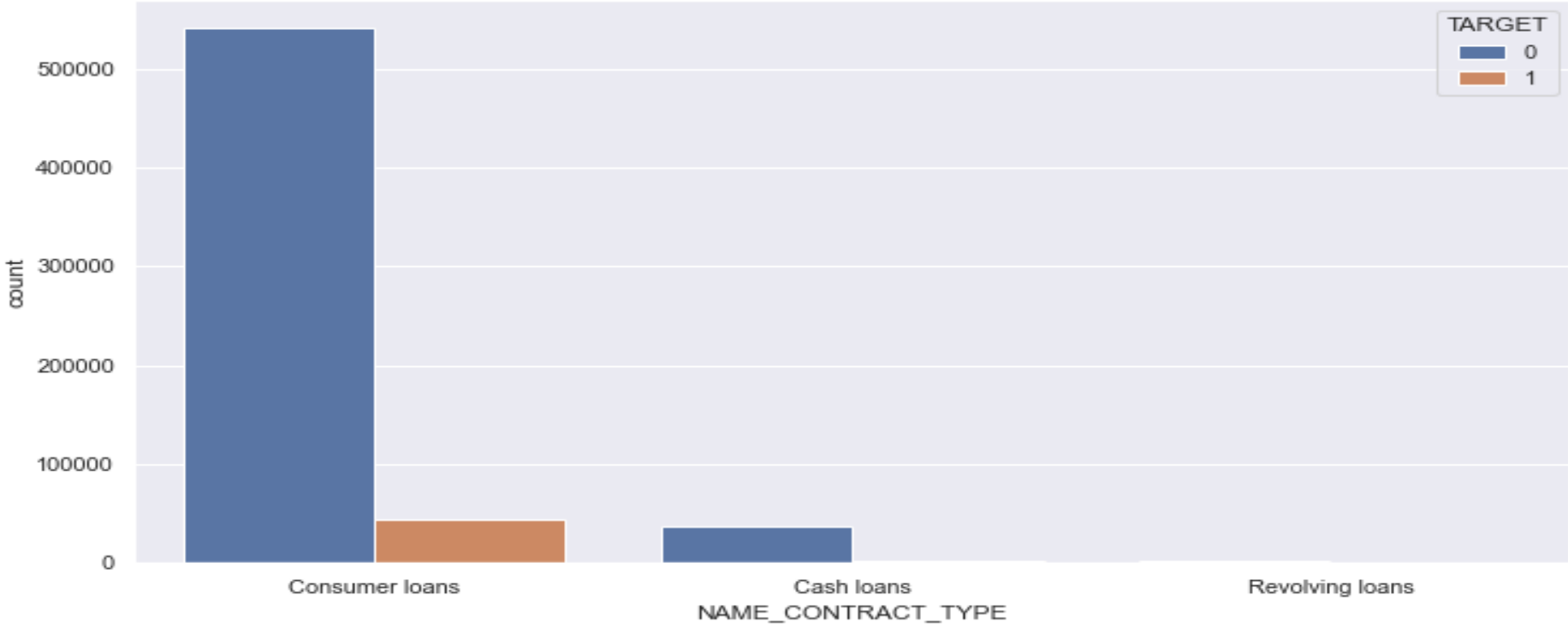
Analysis of Down payment done by Customers:

RATE_DOWN_PAYMENT	
0.50	0.026869
0.70	0.104262
0.90	0.200852
0.95	0.237161
0.99	0.466691

RATE_DOWN_PAYMENT	
0.50	0.091654
0.70	0.108909
0.90	0.212209
0.95	0.289673
0.99	0.507617

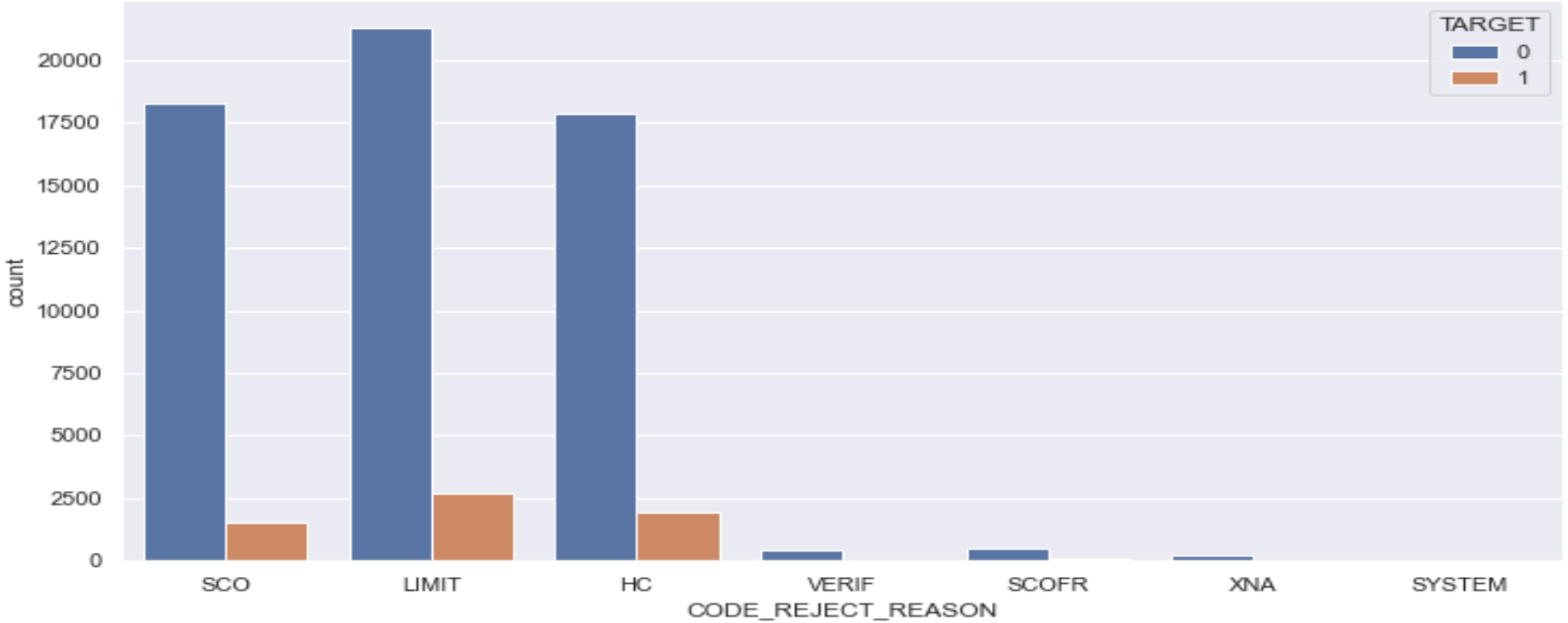
Analysis of Customers intrests in the Previous Application Data

- *Customers are mostly intrested in Consumer loans (as per Previous application data)*



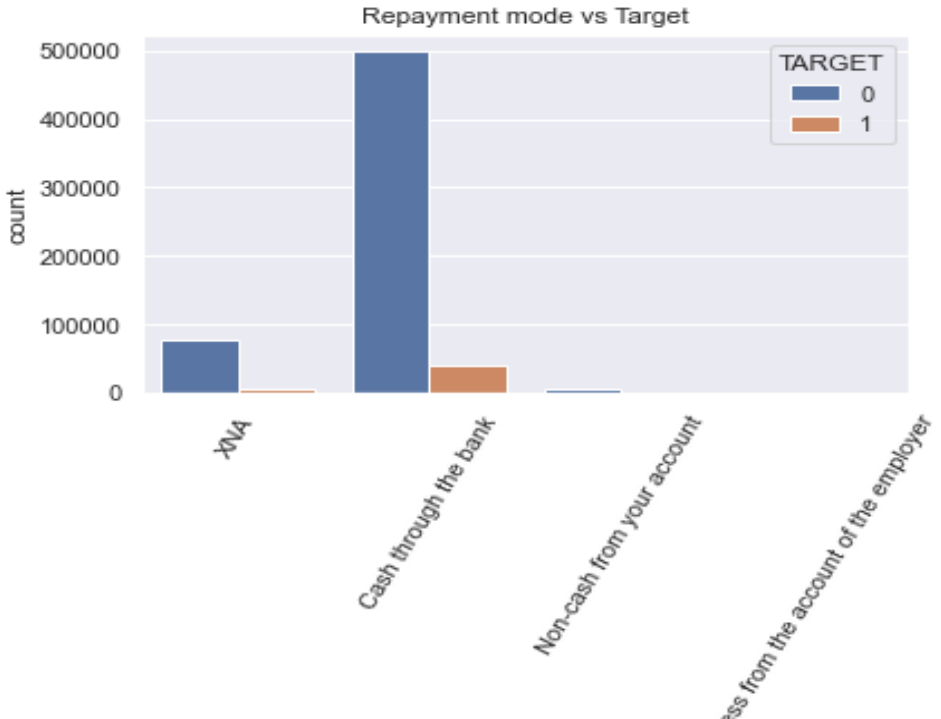
Analysis of why customers previous application got rejected

•*As per plot below :- SCO, LIMIT and HC are the most common reason for rejection of loan application.*



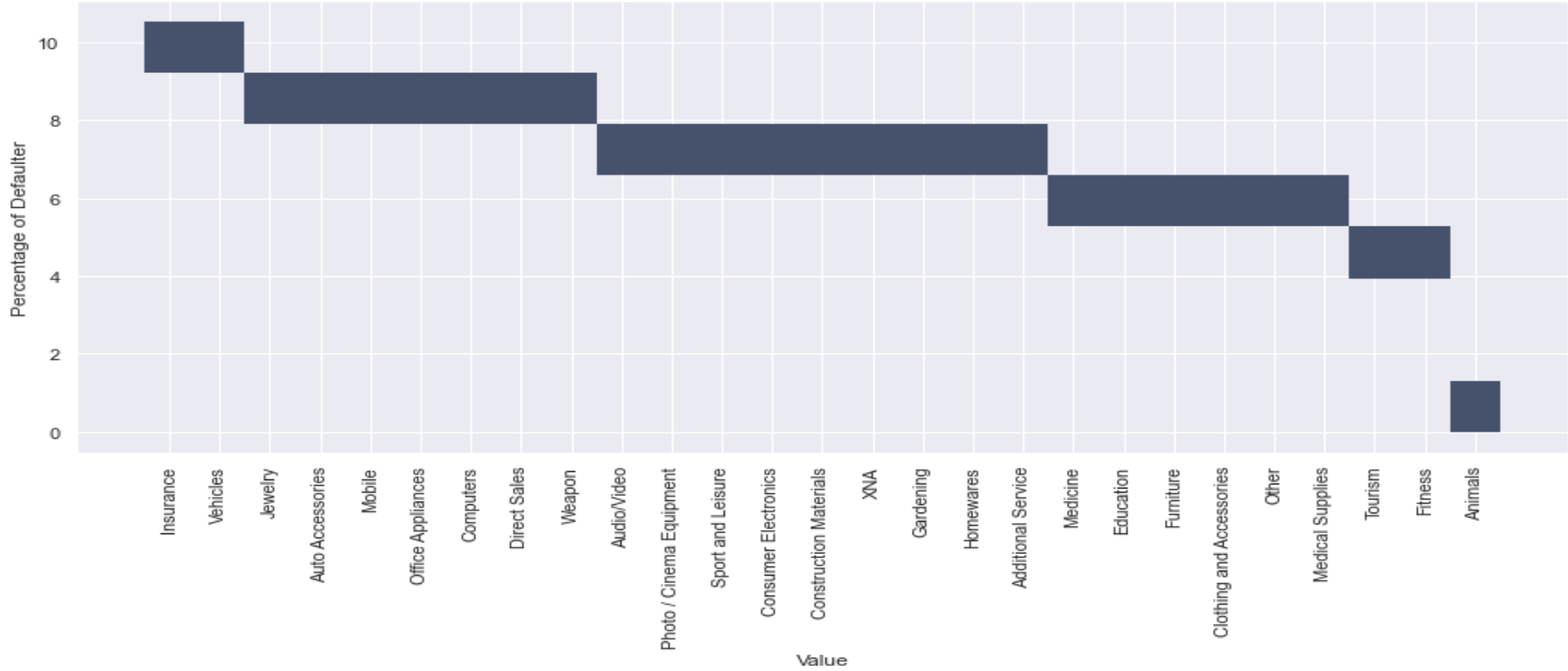
Analysis based on customer type and repayment mode:

- *Most of the customers have taken or applied for loan before i.e. they are mostly repeaters.*
- *Customers mostly prefer to pay the loan back through this medium: Cash through the bank*



Analysis based on category of goods:

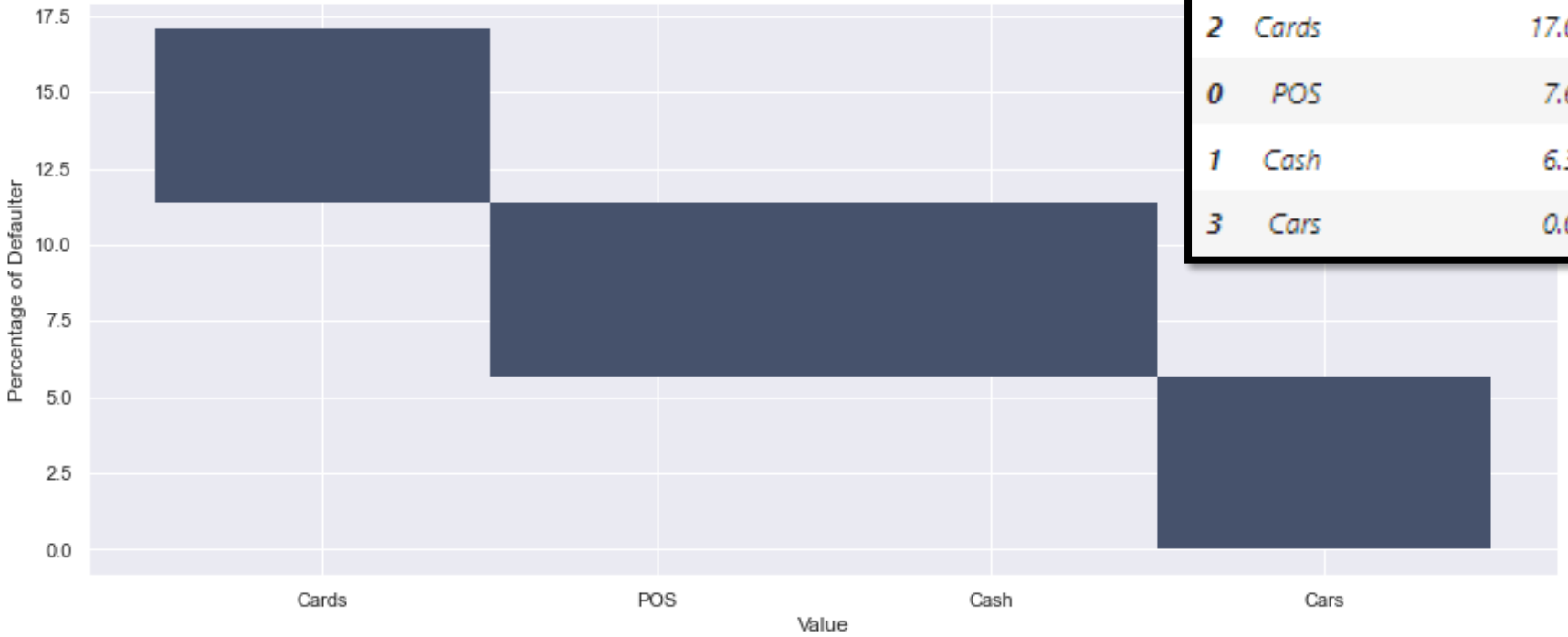
- *Most of default cases are for the customers who previously applied for either **Insurance** or/and **Vehicles**:*



Analysis on basis of previous application reason (for CASH, POS, CAR, etc):

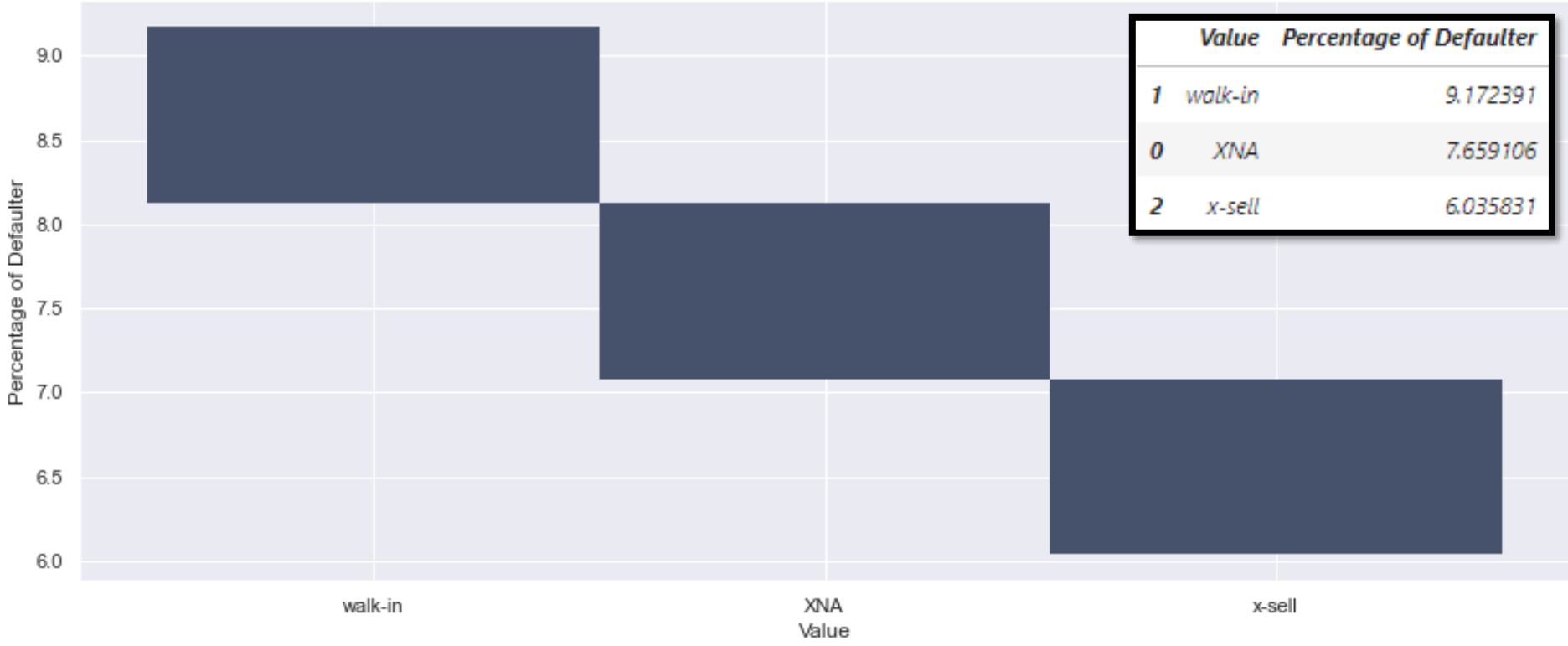
- *For **Cards** defaulter rate is the most*

	Value	Percentage of Defaulter
2	Cards	17.095238
0	POS	7.659301
1	Cash	6.386600
3	Cars	0.000000



Analysis on basis of previous application was x-sell or walk-in:

•*Among both X-sell and Walk-in, the **walk-in** customers defaulted the most in current loan(i.e. 9%).*

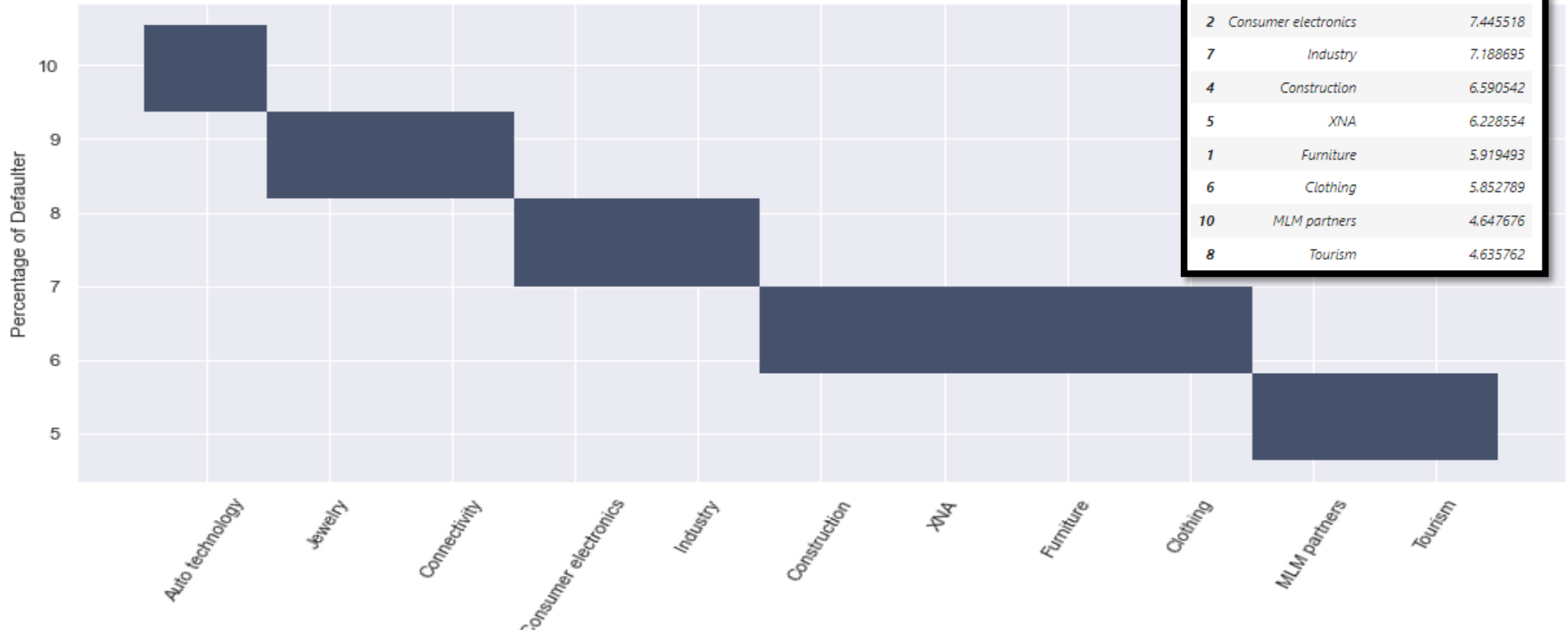


	Value	Percentage of Defaulter
1	walk-in	9.172391
0	XNA	7.659106
2	x-sell	6.035831

Analysis on basis of the industry of the seller:

- *"Auto technology" has highest rate of defaulter customers*
- *"MLM partners" has lowest number of defaulter customers*

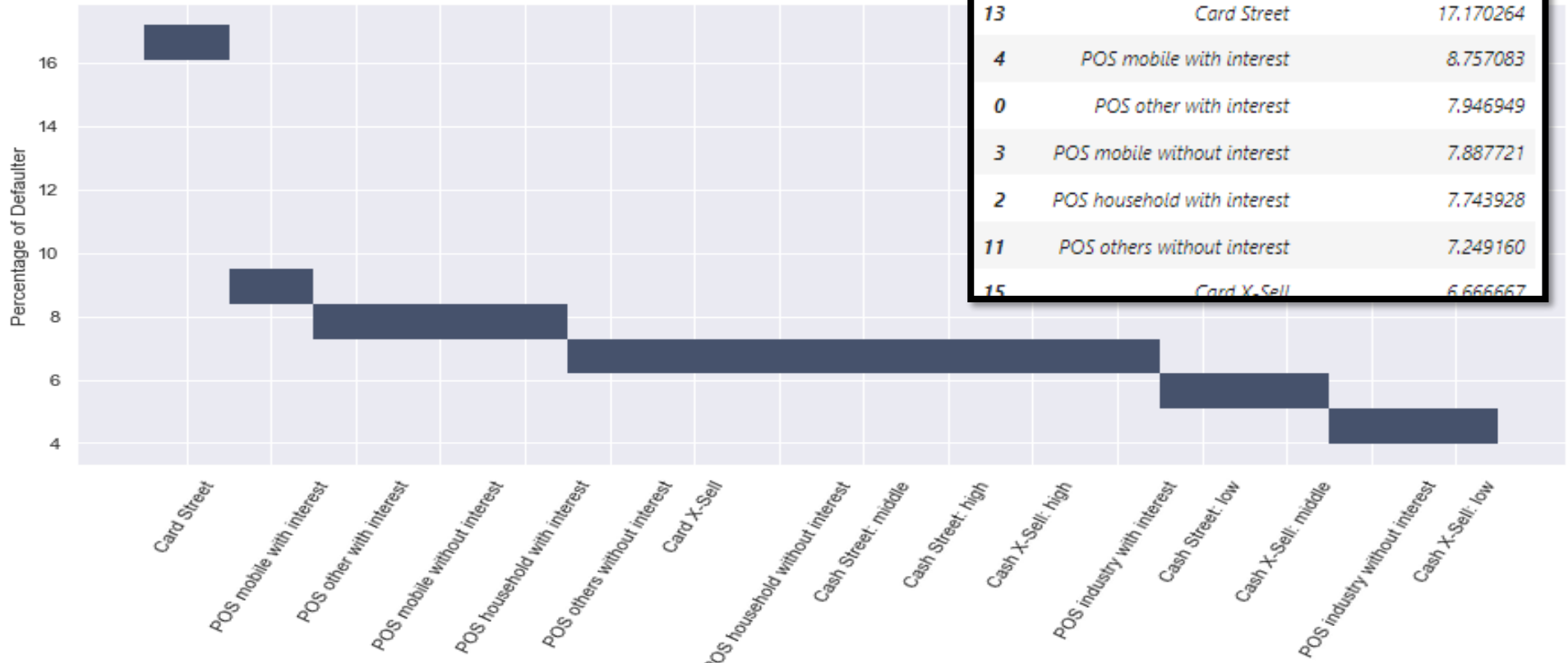
	Value	Percentage of Defaulter
0	Auto technology	10.554371
9	Jewelry	8.997050
3	Connectivity	8.777794
2	Consumer electronics	7.445518
7	Industry	7.188695
4	Construction	6.590542
5	XNA	6.228554
1	Furniture	5.919493
6	Clothing	5.852789
10	MLM partners	4.647676
8	Tourism	4.635762



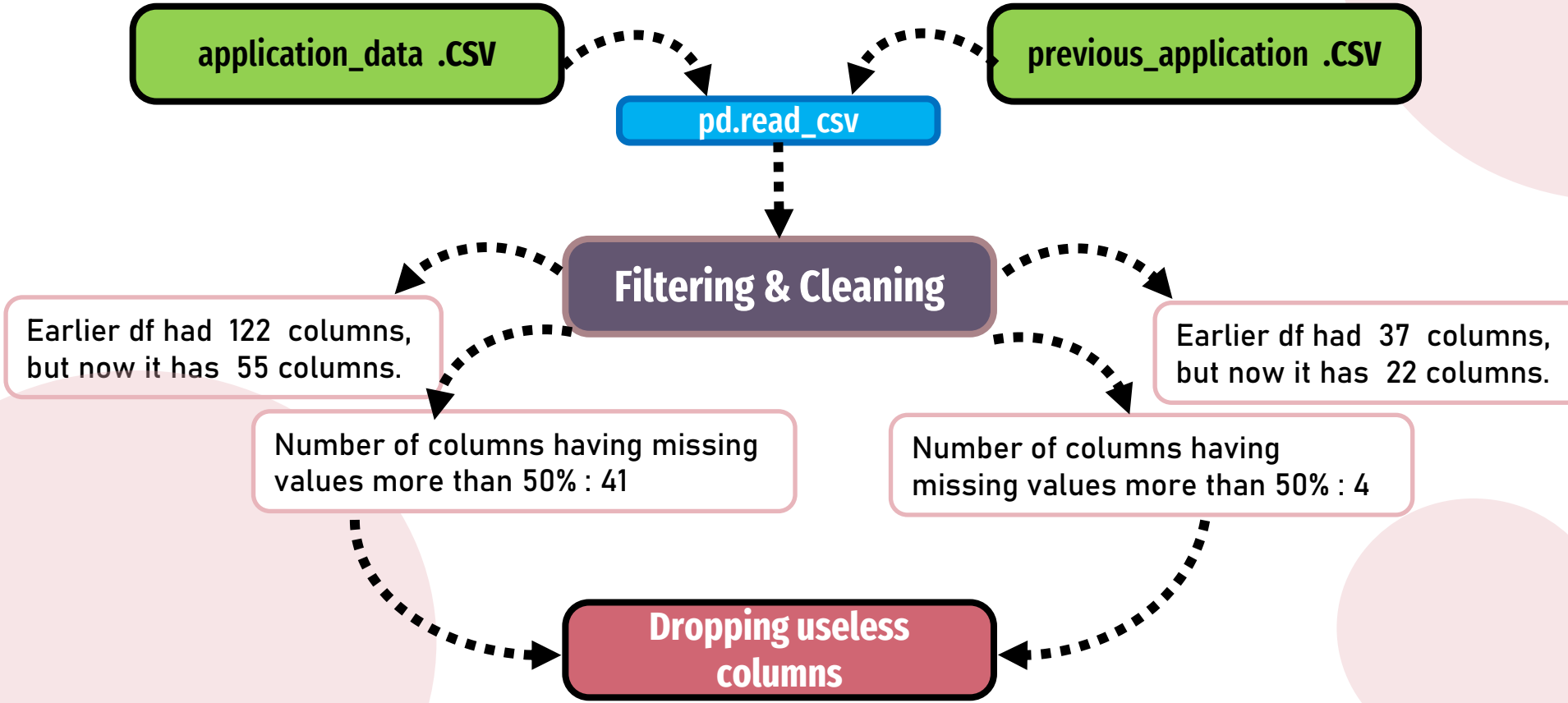
Analysis on basis of the detailed product combination of the previous application

- *"Card Street" has highest rate of defaulter customers*
- *"Cash X-cell low" has lowest number of defaulter customers*

	Value	Percentage of Defaulter
13	Card Street	17.170264
4	POS mobile with interest	8.757083
0	POS other with interest	7.946949
3	POS mobile without interest	7.887721
2	POS household with interest	7.743928
11	POS others without interest	7.249160
15	Card X-Sell	6.666667



Conclusions



Conclusions

Columns used to analyse data:

NAME_CONTRACT_STATUS,
AMT_ANNUITY, AMT_DOWN_PAYMENT,
NAME_CONTRACT_TYPE, CODE_GENDER,
NAME_HOUSING_TYPE,
REG_REGION_NOT_WORK_REGION,
REG_CITY_NOT_LIVE_CITY,
REG_CITY_NOT_WORK_CITY,
REGION_RATING_CLIENT,
REGION_RATING_CLIENT_W_CITY,
FLAG_DOCUMENT_2,
FLAG_DOCUMENT_3,
FLAG_DOCUMENT_4,
FLAG_DOCUMENT_5,
FLAG_DOCUMENT_6,
FLAG_DOCUMENT_7,
FLAG_DOCUMENT_8,
FLAG_DOCUMENT_9,
FLAG_DOCUMENT_10,
FLAG_DOCUMENT_11,
FLAG_DOCUMENT_12,
FLAG_DOCUMENT_13,
FLAG_DOCUMENT_14,
FLAG_DOCUMENT_15,
FLAG_DOCUMENT_16,
FLAG_DOCUMENT_17,
FLAG_DOCUMENT_18,
FLAG_DOCUMENT_19,
FLAG_DOCUMENT_20,
FLAG_DOCUMENT_21,
FLAG_MOBIL,
FLAG_EMP_PHONE,
FLAG_WORK_PHONE,
FLAG_CONT_MOBILE,
FLAG_PHONE,
FLAG_EMAIL,
FLAG_OWN_CAR,
CNT_FAM_MEMBERS,
NAME_FAMILY_STATUS,
NAME_TYPE_SUITE,
NAME_INCOME_TYPE,
NAME_EDUCATION_TYPE,
OCCUPATION_TYPE,
CODE_GENDER,
ORGANIZATION_TYPE,
NAME_CONTRACT_TYPE,
WEEKDAY_APPR_PROCESS_START,
AMT_INCOME_TOTAL,
AMT_GOODS_PRICE,
NAME_CONTRACT_STATUS,
RATE_DOWN_PAYMENT,
CODE_REJECT_REASON,
CODE_REJECT_REASON,
NAME_PAYMENT_TYPE,
NAME_GOODS_CATEGORY,
NAME_PORTFOLIO,
NAME_PRODUCT_TYPE,
NAME_SELLER_INDUSTRY,
PRODUCT_COMBINATION,
FLAG_OWN_REALTY,
OWN_CAR_AGE,
CNT_CHILDREN,

Summary

1. *We can easily observe that **repayment rate** increases with increase in **defaulter rate**.*
2. *Most of the customers live in **House/Apartment***
3. *Most of the defaulters also live in **House/Apartment***
4. *Rented apartments have higher rate of defaulters as they can always shift(as per the data)*
5. *We can easily observe that **defaulter rate** increases when:*
    1. *Current region is same as working region.*
    2. *Current city is same as city in which client is living in.*
    3. *Current city is same as city in which client is working in.*

6. *Most of the customers live in Region rated **2nd** among 3 regions.*
7. *Defaulters are the most in **region 2** then they decrease in **region 3** and least are in **region 1**.*
8. *Increase in defaulters per region is natural because the number of customers also increases.*
9. *We can easily observe that **Most** of the documents have same trend **excluding FLAG_DOCUMENT_3***
10. *Cutomers who have submitted the documents **5,6,8,9,16 and 18** are less likely to default as per the trend.*
11. *We can easily observe that **defaulters** havent provided:*
    1. *Their Email address(es).*
    2. *Their Home phone.*

Summary

12. ***FLAG_PHONE*** *and* ***FLAG_WORK_PHONE*** *are identical hence either can be dropped.*
13. *Defaulters are generally under category* ***Civil Marriage*** *and* ***Single applicants****.*
14. *Most of the defaulters are having* ***1-4*** *family members.*
15. *Most of the defaulters are having* ***0-2*** *childrens.*
16. *Customers having* ***9*** *or* ***11*** *children can be suspected as outliers in the df.*
17. *Customers having* ***11*** *or* ***13*** *members can be suspected as outliers in the df.*
18. *Most of the defaulters were* ***unacompanied*** *during Loan screening process.*
19. *We can easily observe that defaulters typically dont own a* ***car****.*
20. *Also notice that most of the customers own a* ***realty*** *as per the trend.*
21. *People not having a* ***realty*** *and* ***car*** *and have higher chances of default than the people who own* ***realty*** *and* ***car***
22. *Defaulter or not, most applicants have car age between 0-25 years.*
23. *Defaulters are generally* ***Unemployed*** *or in* ***Maternity leave****.*
24. *Most of the customers are* ***Working****.*
25. *Most of the customers have not completed their* ***Higher*** *education.*
26. *Most of the working defaulters are working as* ***Low-skilled labourers****.*
27. *The customers are having age between* ***20*** *to* ***69****.*
28. *People of age* ***25-35*** *have higher default rate.*
29. *Default cases are less for applicants more than* ***40 years*** *old..*
30. *The customers are mostly opting for* ***Cash Loans****.*
31. *The column* ***WEEKDAY_APPR_PROCESS_START*** *is not really helpful to derive any metrics.*

Summary

32. *The columns **AMT_ANNUITY** and **AMT_ANNUITY** have some outliers.*
33. *The columns **AMT_ANNUITY** and **AMT_ANNUITY** have similar trends.*
34. *The applicants whose previous loans were approved are more likely to pay current loan in time.*
35. *The customers who **paid less** of down payment in previous application have higher cases of default.*
36. *Customers are mostly intrested in **Consumer loans** (as per Previous application data)*
37. *As per plot below :- SCO, LIMIT and HC are the most common reason for rejection of loan application.*
38. *Most of default cases are for the customers who previously applied for either **Insurance** or/and **Vehicles***
39. *Among both X-sell and Walk-in, the **walk-in** customers defaulted the most in current loan(i.e. 9%).*
40. *"Auto technology" has highest rate of defaulter customers*
41. *"MLM partners" has lowest number of defaulter customers*
42. *"Card Street" has highest rate of defaulter customers*
43. *"Cash X-cell low" has lowest number of defaulter customers*

Thanks !

-Vivek Kumar