Exploratory Data Analysis

Loan Risk analysis based on
available dataset by EDA

EDA Analysis by :

- Debopriya Ghosh

BUSINESS OBJECTIVE

To approve loan application of clients who are capable of repaying the loans.
In other words, the company wants to understand the driving factors 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.

EDA on available Current Application dataset and Previous Application dataset to understand how customer attributes and loan attributes influence the tendency of default.

Identifying 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 at a higher interest rate to risky applicants.

TYPES OF DECISSION ON LOAN APPLICATION

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.

RISKS ASSOCIATED WITH DECISSION

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.

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.

Using EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected and those who are unlikely to pay are not approved.

DATASET

Application Dataset:

- **122** Columns
- **307511** rows of unique Application ID

Previous Application Dataset:

- **37** Columns
- **1670214** rows of unique Previous Application ID

Significant Categorical variables – For Approving or Rejecting Client's Loan:

Below are the categorical variable that to be considers for making decisions on a new client Application data Variable:

- NAME_CONTRACT_TYPE
- NAME_EDUCATION_TYPE
- NAME_HOUSING_TYPE
- NAME_TYPE_SUITE
- NAME_FAMILY STATUS
- OCCUPATION_TYPE
- NAME_INCOME_TYPE
- FLAG_OWN_CAR
- FLAG_OWN_REALTY

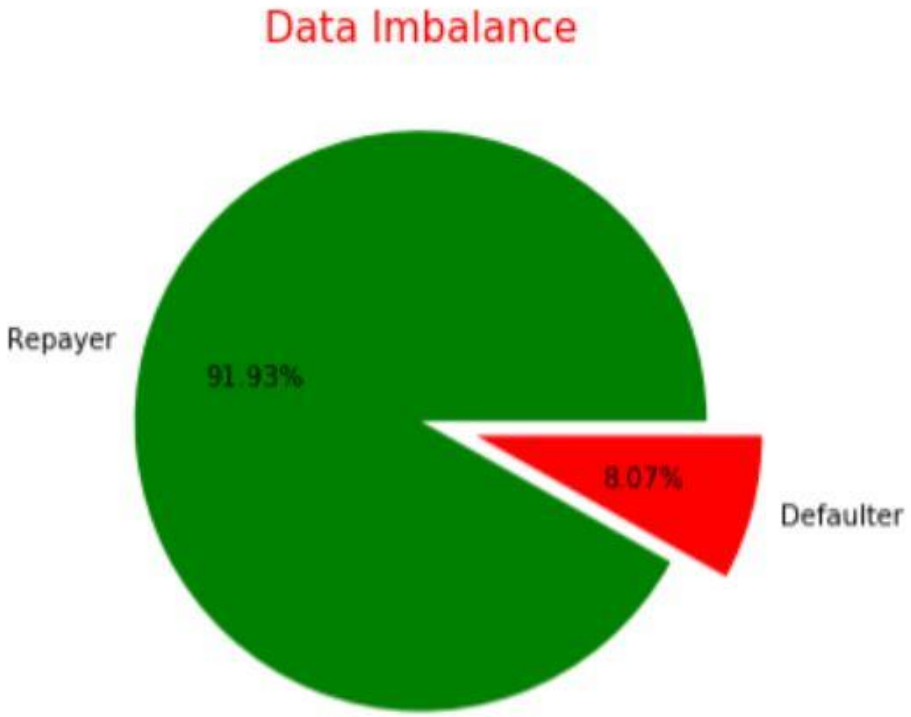
Merged data variable : Previous data has lot of influence on the decision. Below categorical variable are more significant than other for making decisions.

- NAME_PORTFOLIO
- NAME_TYPE_SUITE_PREV
- NAME_PRODUCT_TYPE
- NAME_SELLER_INDUSTRY
- NAME_INCOME_TYPE

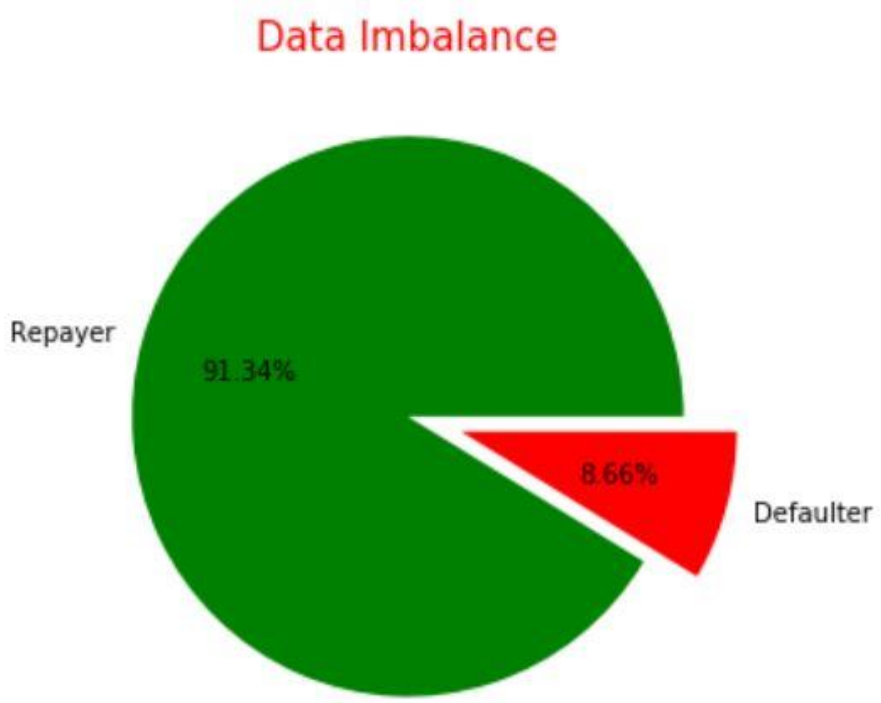
The insight and the result of the analysis of each variable are mentioned in upcoming slides.

Univariate Analysis Segment over TARGET variable 0 for no payment difficulties and 1 for defaulters:

- Application Data
- Application and Previous Merged Data



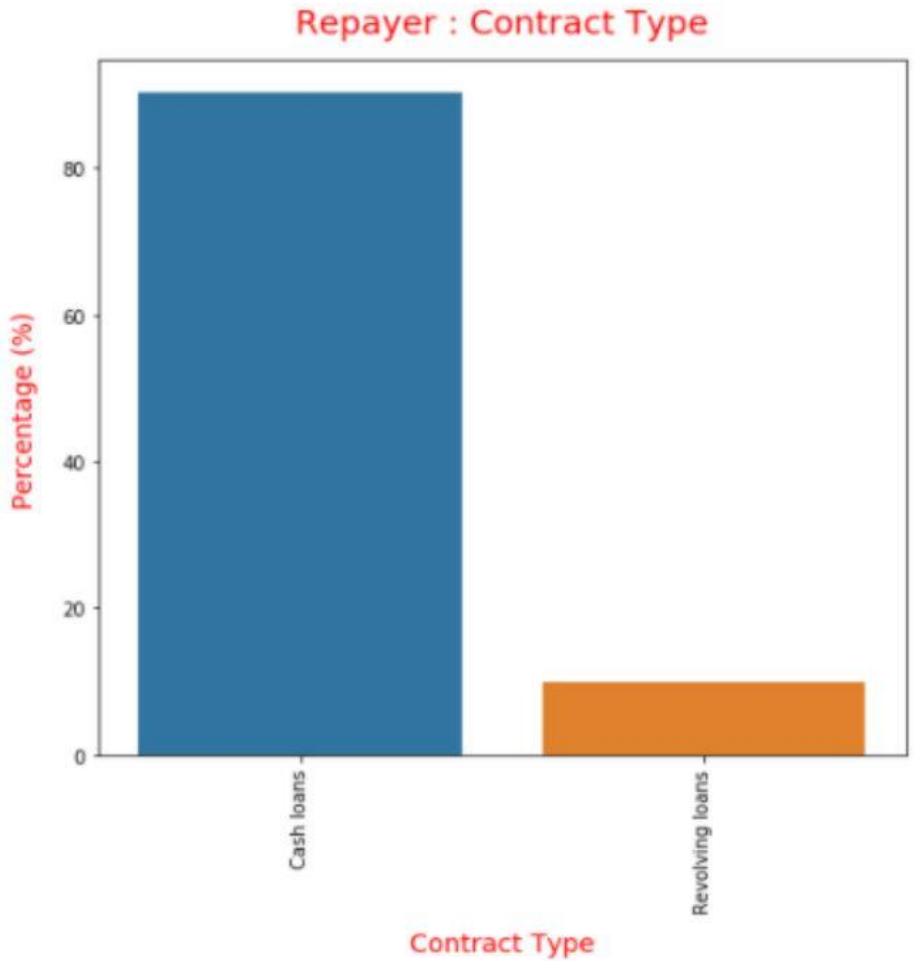
Imbalance of the data in application data Frame for TARGET 0 and 1

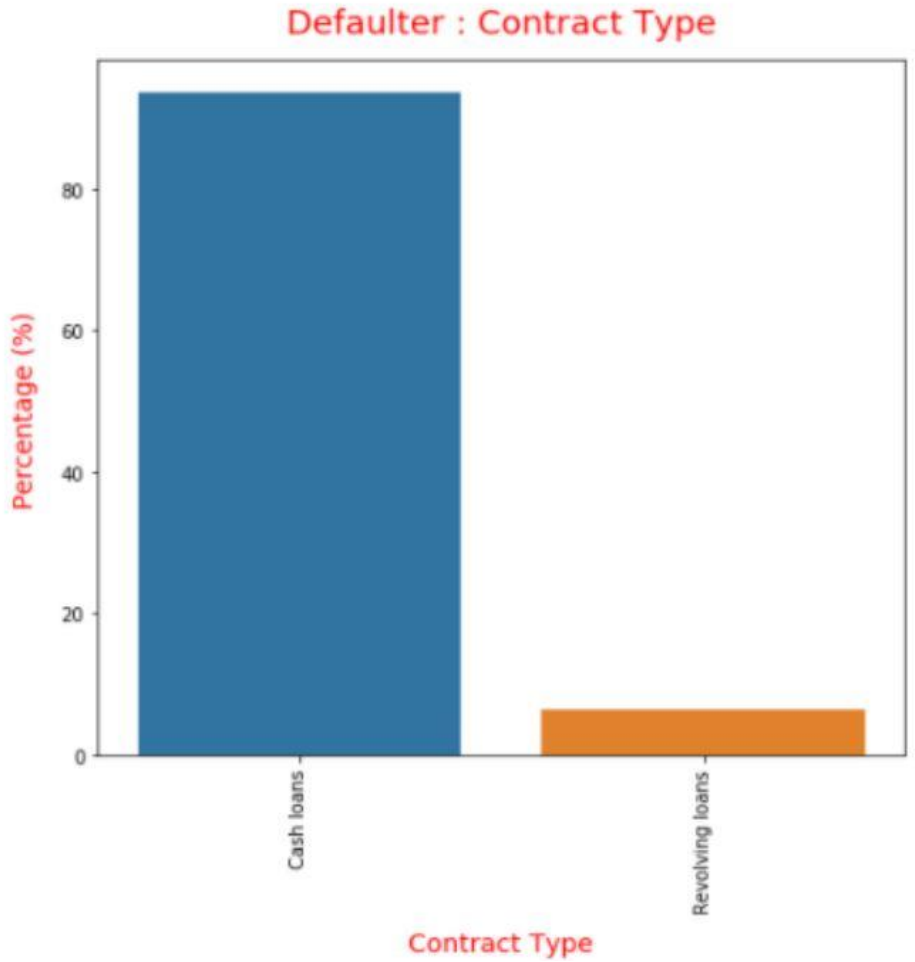


Imbalance of the data in merged dataset of application data and previous data for TARGET 0 and 1

Univariate Analysis Segment over TARGET variable 0 for no payment difficulties and 1 for defaulters on application data frame:

- CONTRACT_TYPE : Cash and Revolving Loan



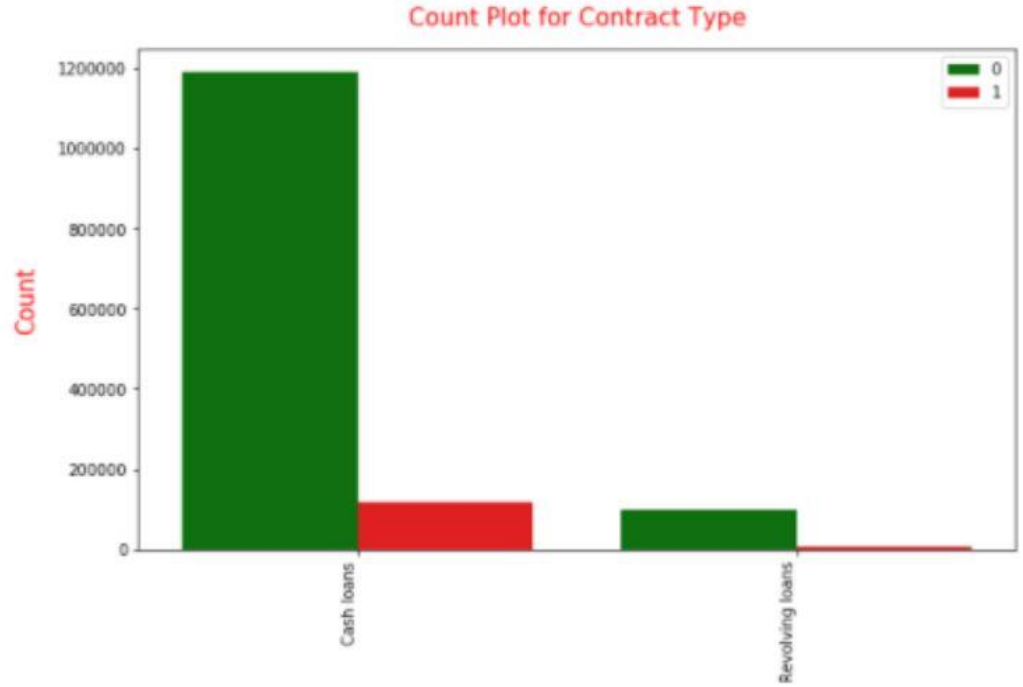


Customer taking the cash loan is most likely to make payments on time. Whereas Customer taking revolving loans is most likely to make payments on time.

Univariate Analysis Segment over TARGET variable 0 for no payment difficulties and 1 for defaulters on merged dataset (previous and application data frame):

- CONTRACT_TYPE : Consumer , Cash and Revolving Loan

NAME_CONTRACT_TYPE_PREV	TARGET	Counts	Percentage
Cash loans	0	569567	90.87%
	1	57197	9.13%
Consumer loans	0	577049	92.29%
	1	48207	7.71%
Revolving loans	0	144475	89.53%
	1	16893	10.47%
XNA	0	250	79.87%
	1	63	20.13%



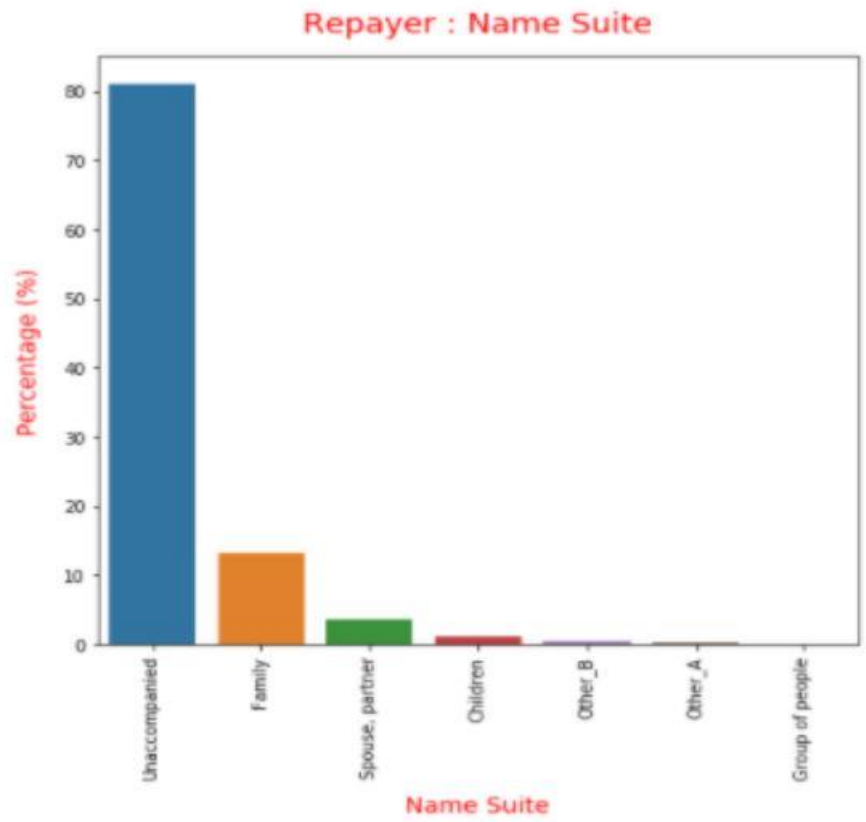
NAME_CONTRACT_TYPE	TARGET	Counts	Percentage
Cash loans	0	1190586	91.09%
	1	116529	8.91%
Revolving loans	0	100755	94.53%
	1	5831	5.47%

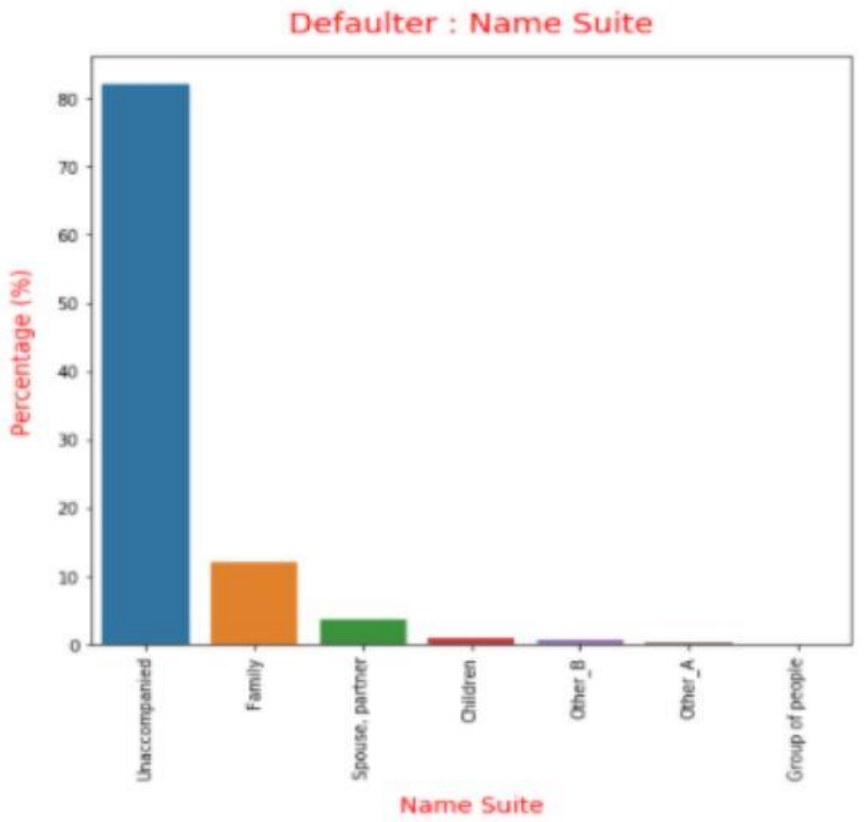
- In Cash Loans - 9.13% , In Consumer loans - 7.71% , In Revolving loans - 10.47%

Hence, we can say that in previous dataset customers who took Revolving Loan was most likely having payment difficulties.

Univariate Analysis Segment over TARGET variable 0 for no payment difficulties and 1 for defaulters on application data frame :

- NAME_TYPE_SUITE



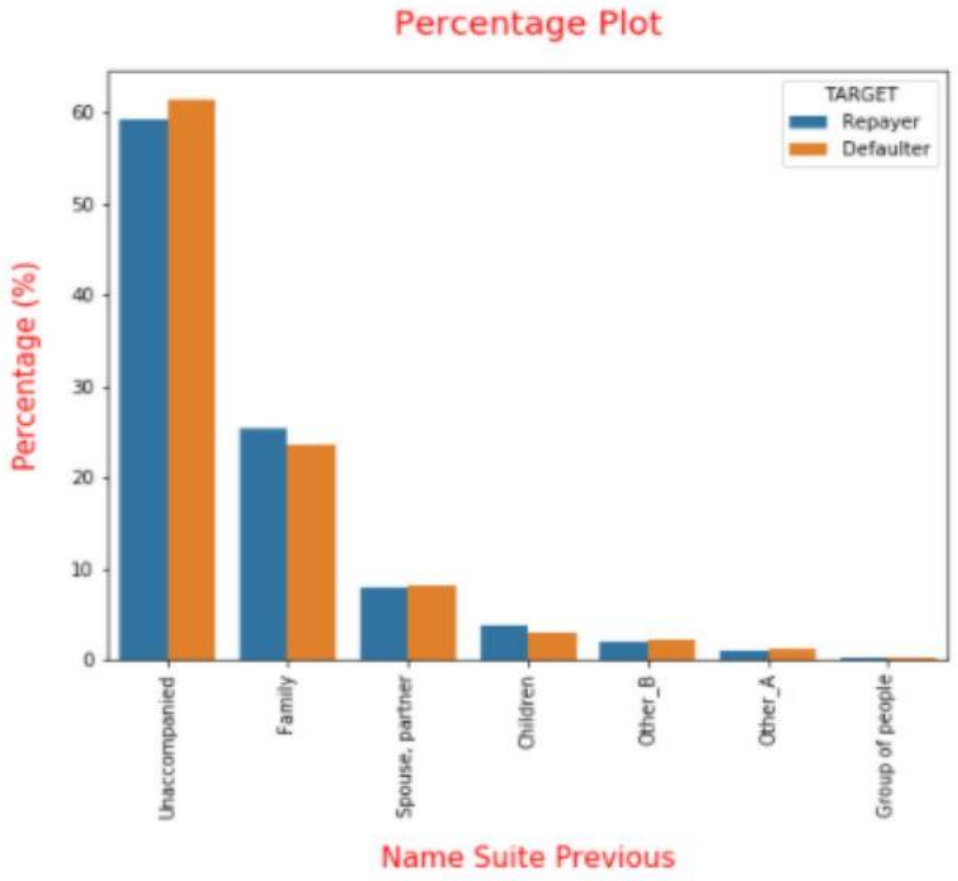


Customer who are unaccompanied during loan application is most likely to make payments on time.

- From the above univariate percentage plot, we can observe that more percentage (> 80%) of defaulters also are from customer who are unaccompanied during loan application .**Hence this group could be very risky for bank to give loans as `defaulter` percentage are more than re payers.**

Univariate Analysis Segment over TARGET variable 0 for no payment difficulties and 1 for defaulters on merged dataset (previous and application data frame):

- NAME_TYPE_SUITE_PREV

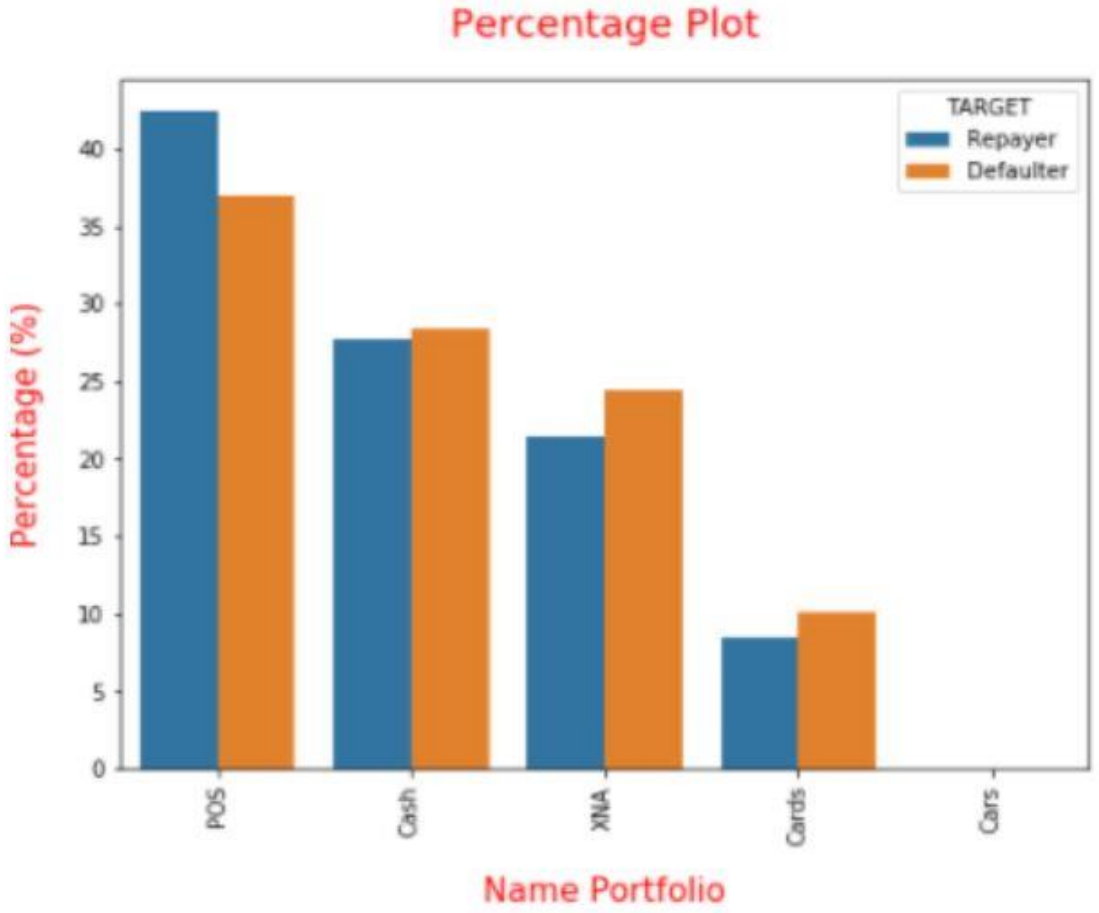


TARGET	NAME_TYPE_SUITE_PREV	Counts	Percentage
0	Unaccompanied	392208	59.3%
	Family	168286	25.44%
	Spouse, partner	52850	7.99%
	Children	25383	3.84%
	Other_B	13793	2.09%
	Other_A	7139	1.08%
	Group of people	1758	0.27%
1	Unaccompanied	35427	61.49%
	Family	13594	23.6%
	Spouse, partner	4736	8.22%
	Children	1723	2.99%
	Other_B	1280	2.22%
	Other_A	681	1.18%
	Group of people	171	0.3%

- It is observed that Unaccompanied customers are most of the defaulters (61.49%) in previous NAME_TYPE_SUITE column, like current application data set.

Univariate Analysis Segment over TARGET variable 0 for no payment difficulties and 1 for defaulters on merged dataset (previous and application data frame):

- NAME_PORTFOLIO

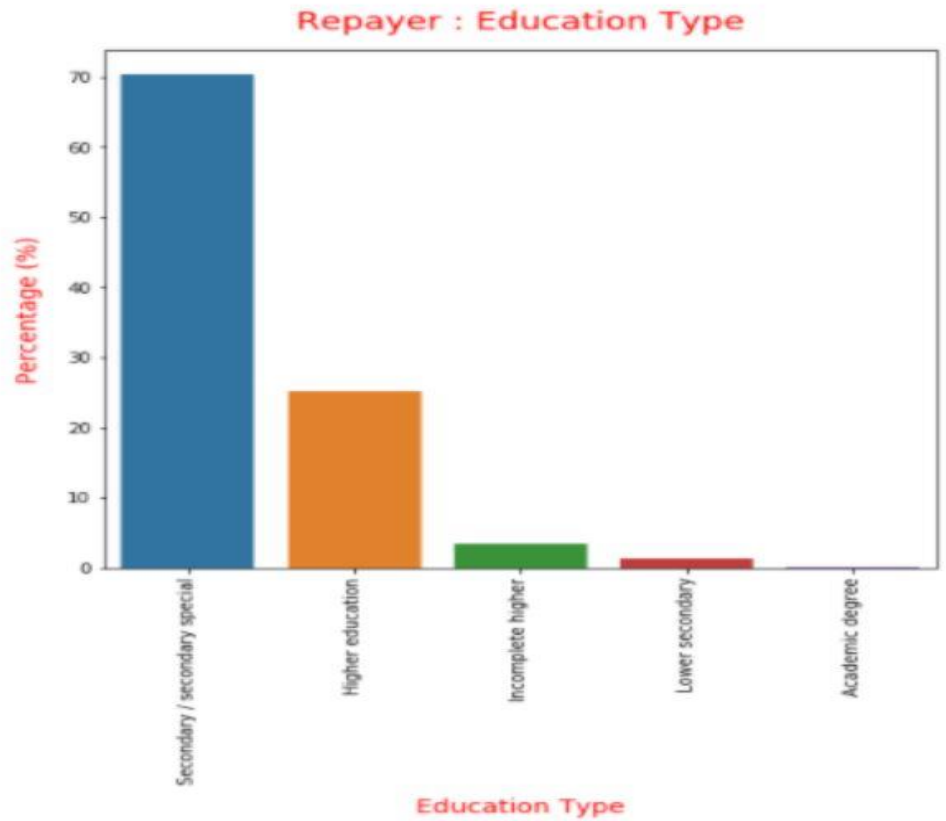


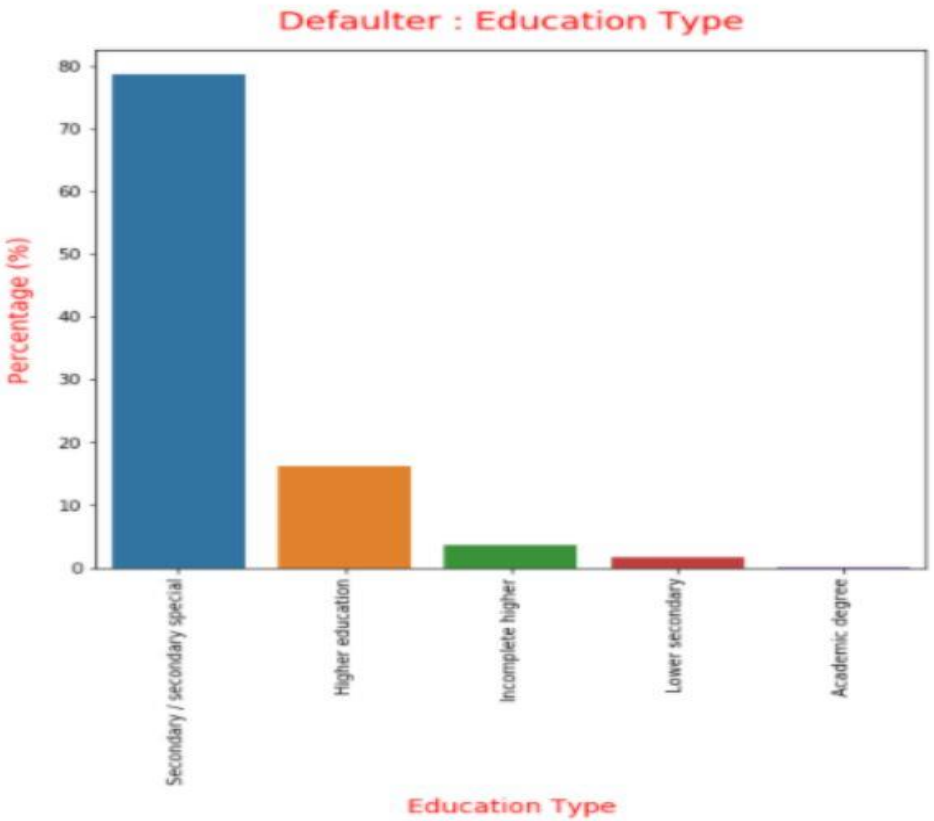
TARGET	NAME_PORTFOLIO	Counts	Percentage
0	POS	547220	42.38%
	Cash	356897	27.64%
	XNA	277276	21.47%
	Cards	109589	8.49%
	Cars	359	0.03%
1	POS	45240	36.97%
	Cash	34766	28.41%
	XNA	29937	24.47%
	Cards	12396	10.13%
	Cars	21	0.02%

- POS segment is to be better towards re payer, while XNA or Unknown category and Cards towards defaulter.

Univariate Analysis Segment over TARGET variable 0 for no payment difficulties and 1 for defaulters on application data frame :

- NAME_EDUCATION_TYPE

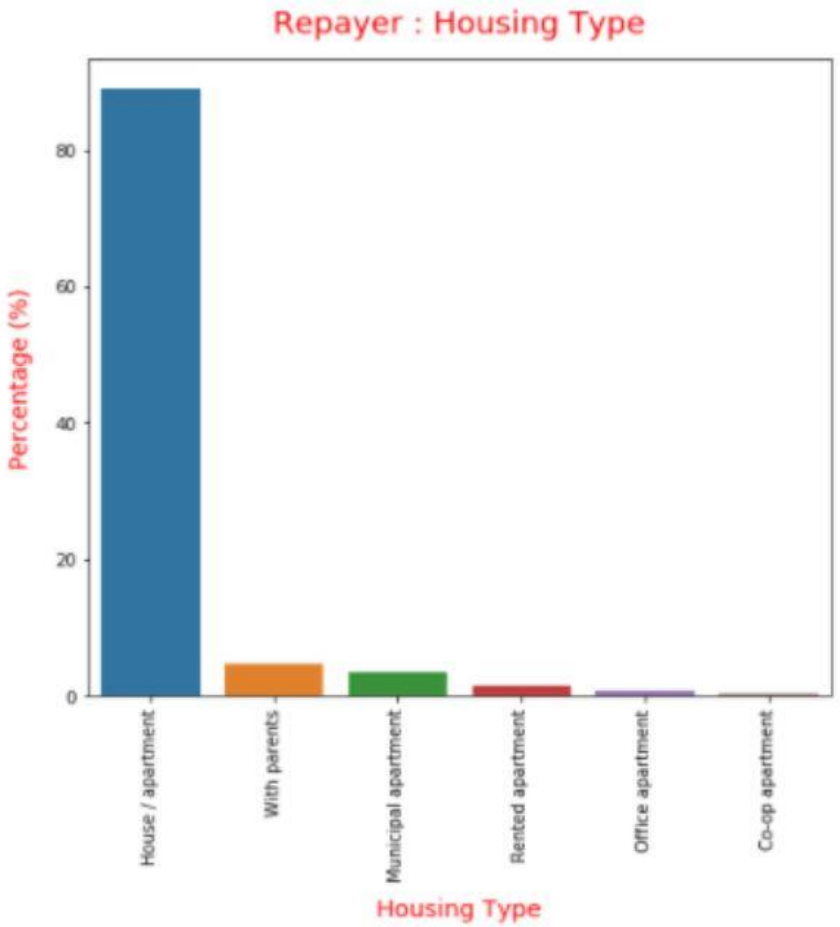


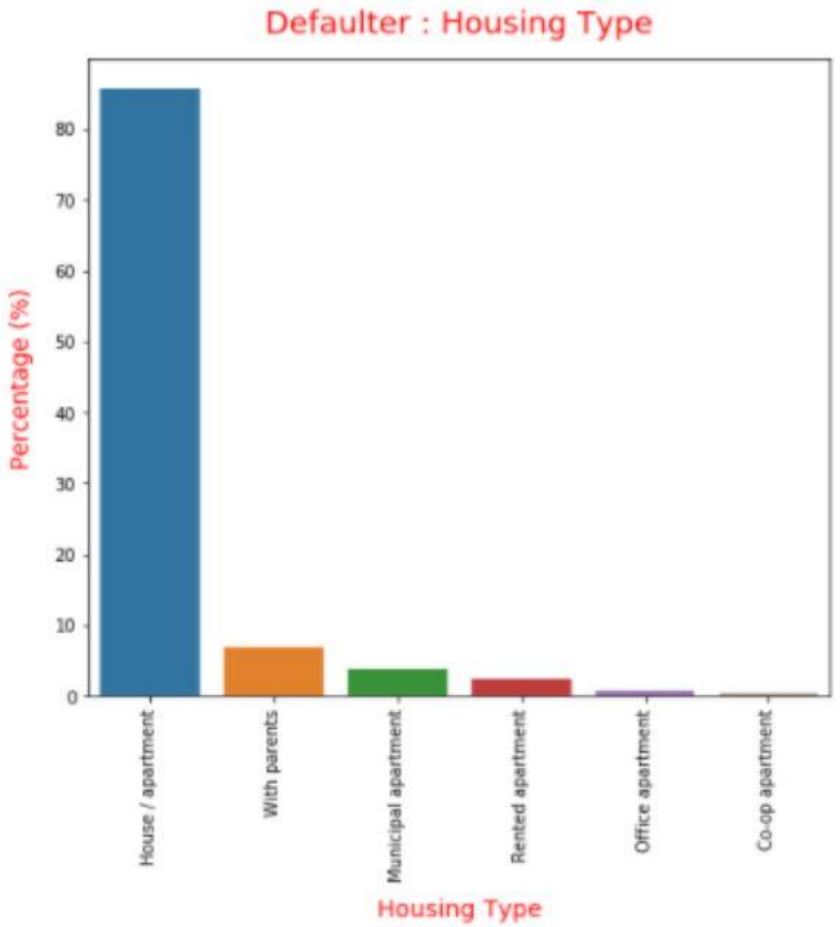


- Customer **with Secondary/Secondary Special education type is the one who most likely to make payments on time**.
- From the above univariate percentage plot, we can observe that more percentage (almost 80%) of defaulters also are from Secondary education type .**Hence this group could be risky for bank to give loans as re payers can turn into defaulter.**
- Whereas Customer with Higher education Degree holders are the one who are also most likely to make payments on time more than 25% are re payers and 15% are defaulters .Hence this group could be more safer for bank to give loans.

Univariate Analysis Segment over TARGET variable 0 for no payment difficulties and 1 for defaulters on application data frame :

- NAME_HOUSING_TYPE

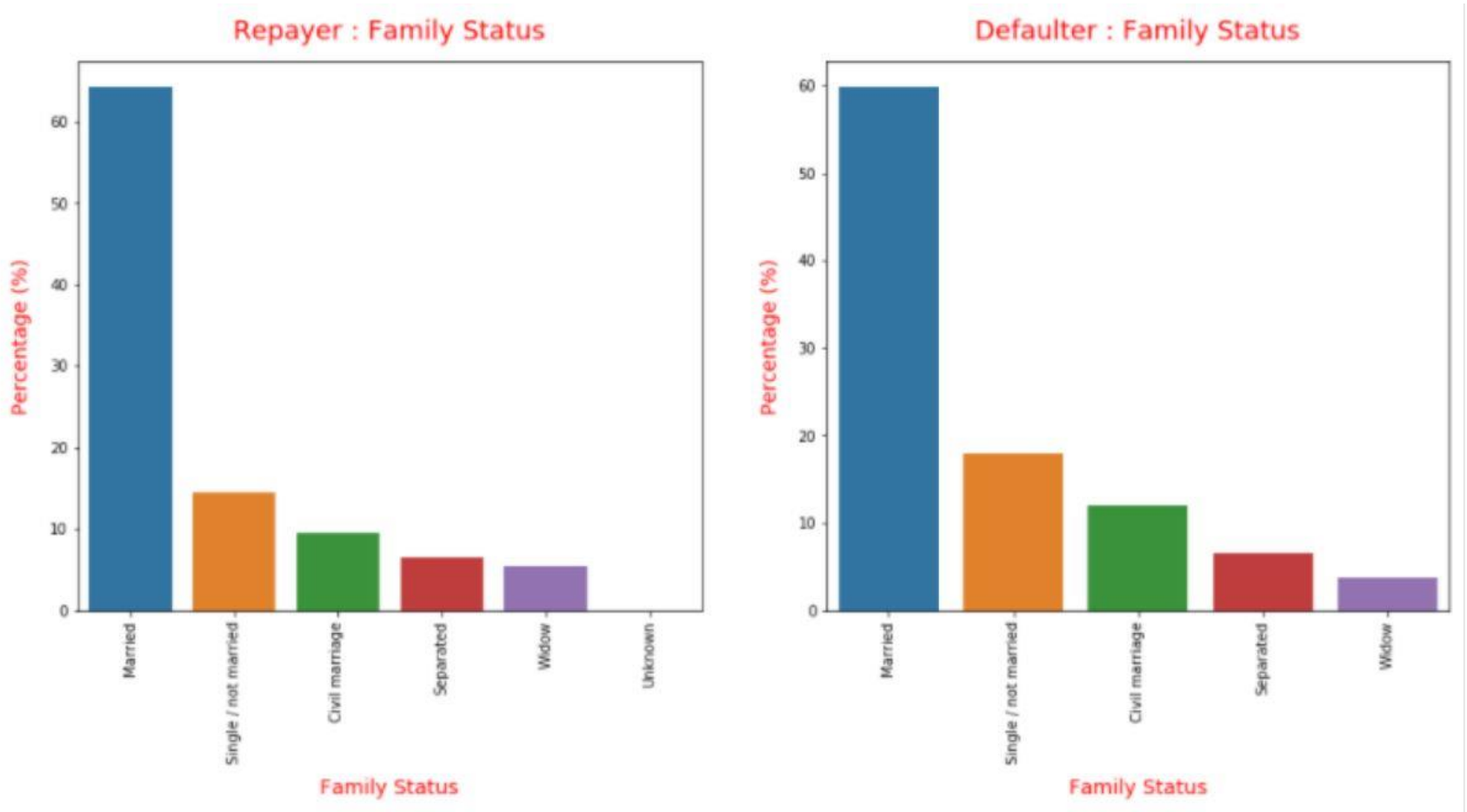




- **Customer owning House/Appartment is most likely to make payments on time.**
- We can also observe that more percentage (almost 85%) of defaulters also are from customer owning House/Appartment. **Hence this group could be very risky for bank to give loans as re-payers and defaulter` are almost similar percentage.**
- Customer living with parents is less likely to make payments on time`. As they have more number of defaulters percentage

Univariate Analysis Segment over TARGET variable 0 for no payment difficulties and 1 for defaulters on application data frame :

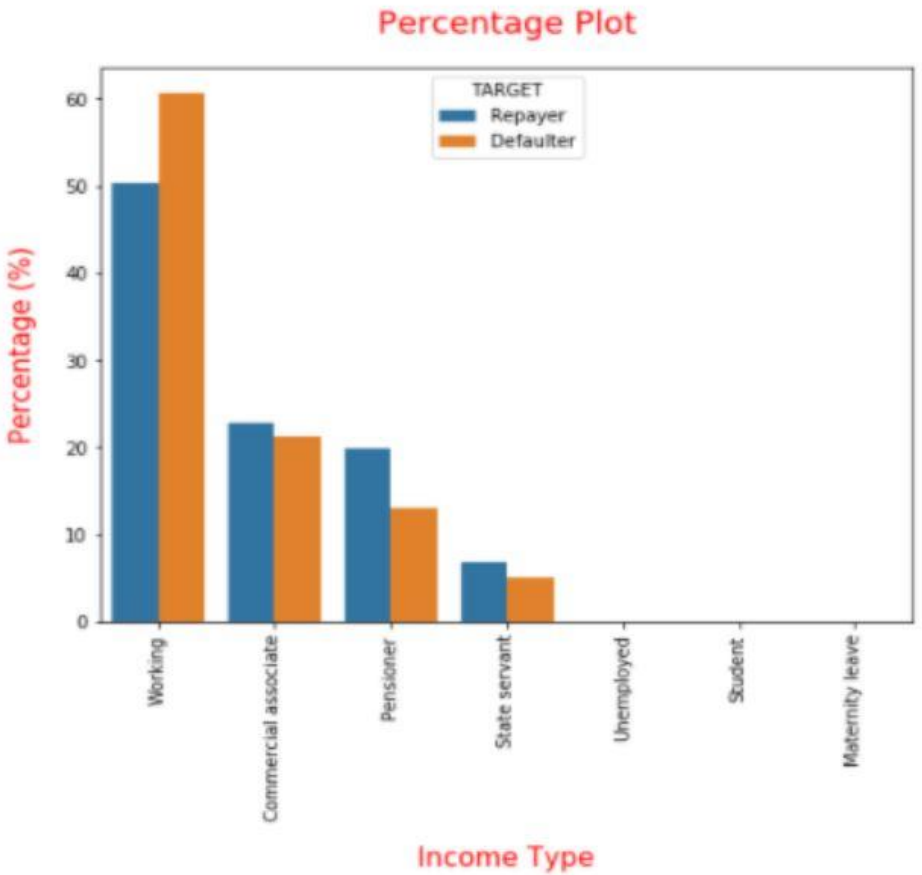
- NAME_FAMILY_STATUS



- **Customer with family status as Married is the one who most likely to make payments on time.**
- From the above univariate percentage plot, we can observe that more percentage (almost 60%) of defaulters also are from family status as Married. **Hence this group could be risky for bank to give loans as re-payers can turn into defaulter**
- Whereas **Customer with family status as Single/not married are the one who less likely to make payments on time. Almost 20% of them are defaulters which we can see from univariate percentage plot.**

Univariate Analysis Segment over TARGET variable 0 for no payment difficulties and 1 for defaulters on merged dataset (previous and application data frame):

- NAME_INCOME_TYPE

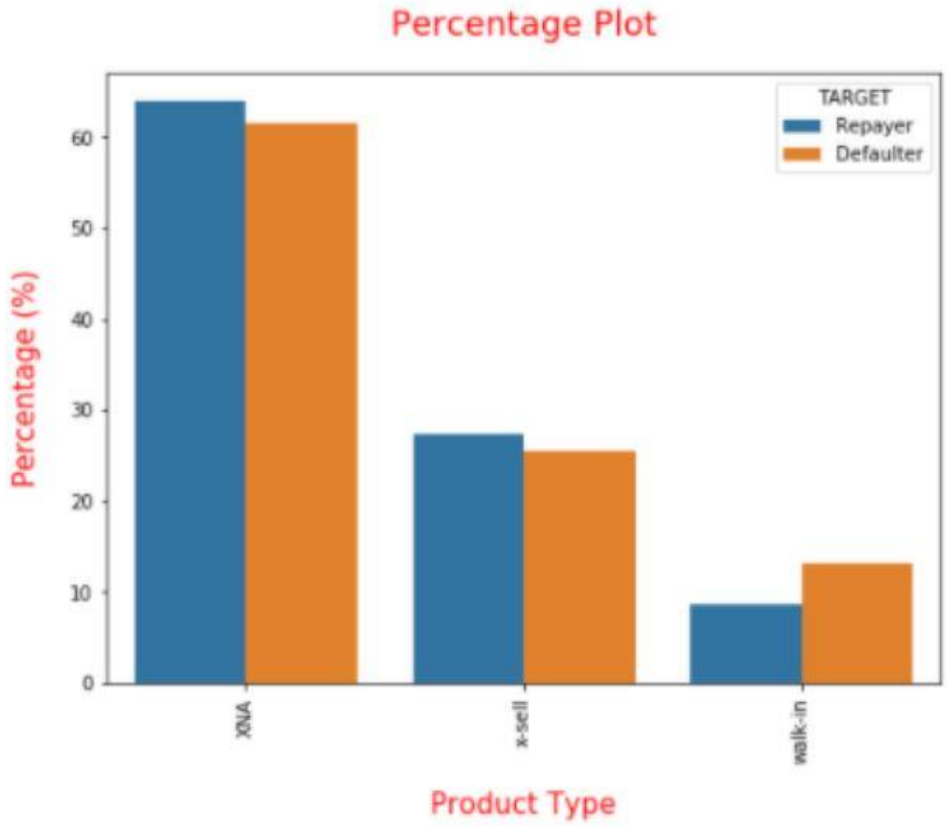


TARGET	NAME_INCOME_TYPE	Counts	Percentage
0	Working	649909	50.33%
	Commercial associate	295396	22.88%
	Pensioner	257586	19.95%
	State servant	88370	6.84%
	Unemployed	56	0.0%
	Student	24	0.0%
1	Working	74180	60.62%
	Commercial associate	26085	21.32%
	Pensioner	15958	13.04%
	State servant	6054	4.95%
	Unemployed	67	0.05%
	Maternity leave	16	0.01%

- Although working professionals **are more likely to repay loan on time among re-payers, it is the most risky category as defaulter turn up is high.**
- **Commercial Assoc** - Slightly lower default turnup .
- **Pensioner** - Lower defaults observed, Hence, could be taken into consideration.
- **State Servant** - This is having lesser defaults.

Univariate Analysis Segment over TARGET variable 0 for no payment difficulties and 1 for defaulters on merged dataset (previous and application data frame):

- NAME_PRODUCT_TYPE

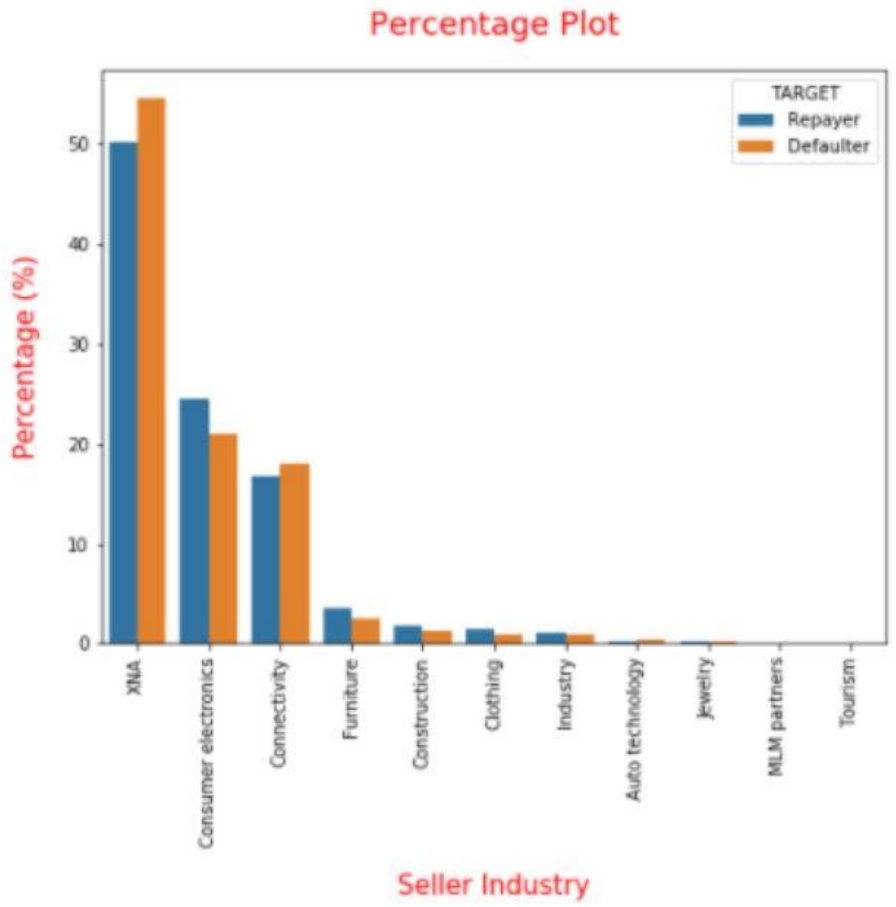


TARGET	NAME_PRODUCT_TYPE	Counts	Percentage
0	XNA	824855	63.88%
	x-sell	354224	27.43%
	walk-in	112262	8.69%
1	XNA	75198	61.46%
	x-sell	31170	25.47%
	walk-in	15992	13.07%

Walk-in type is observed to be high at turning into defaulter compared to XNA/Unknown or x-sell type, which are more of repaying type.

Univariate Analysis Segment over TARGET variable 0 for no payment difficulties and 1 for defaulters on merged dataset (previous and application data frame):

- NAME_SELLER_INDUSTRY



TARGET	NAME_SELLER_INDUSTRY	Counts	Percentage
0	XNA	648400	50.21%
	Consumer electronics	315891	24.46%
	Connectivity	216405	16.76%
	Furniture	45971	3.56%
	Construction	23746	1.84%
	Clothing	18757	1.45%
	Industry	15372	1.19%
	Auto technology	3657	0.28%
	Jewelry	1920	0.15%
	MLM partners	864	0.07%
	Tourism	358	0.03%
1	XNA	66922	54.69%
	Consumer electronics	25810	21.09%
	Connectivity	21945	17.93%
	Furniture	2995	2.45%
	Construction	1671	1.37%
	Industry	1195	0.98%
	Clothing	1139	0.93%
	Auto technology	423	0.35%
	Jewelry	187	0.15%
	MLM partners	57	0.05%
	Tourism	16	0.01%

Consumer electronics is better at repaying, where as, XNA (here XNA is Unknown industry) is a major defaulter turning category.

Significant Numerical variables – For Approving or Rejecting Client's Loan:

Below are the numerical variable that to be considers for making decisions on a new client Application data Variable:

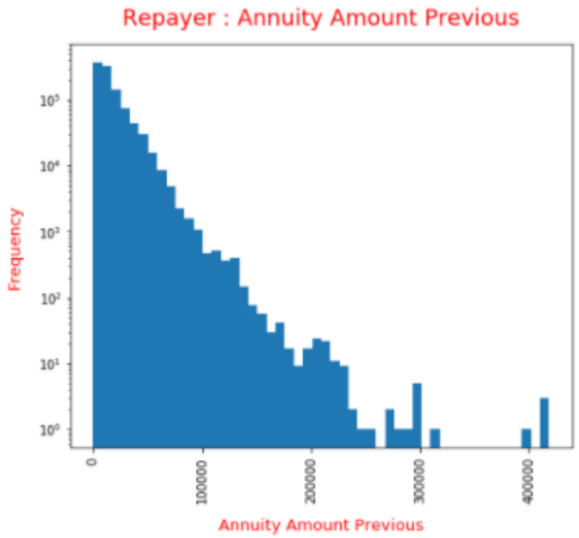
- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE

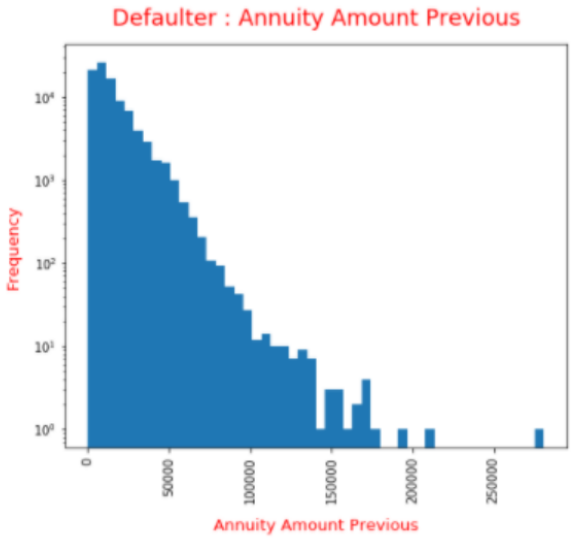
In the previous and current applications following variables are highly correlated.

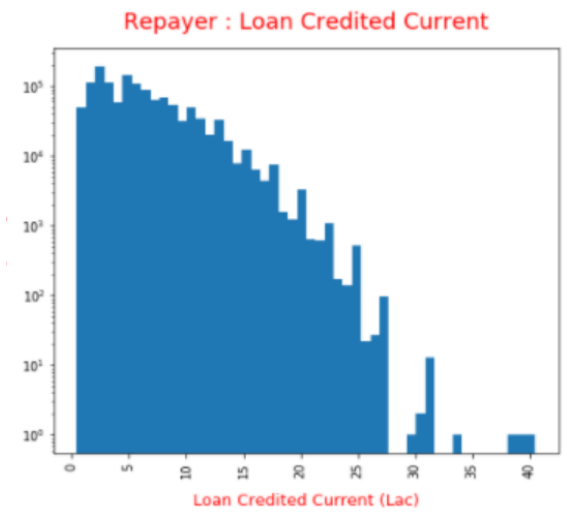
- AMT_ANNUITY_PREV
- AMT_APPLICATION
- AMT_GOODS_PRICE_PREV
- AMT_CREDIT_PREV

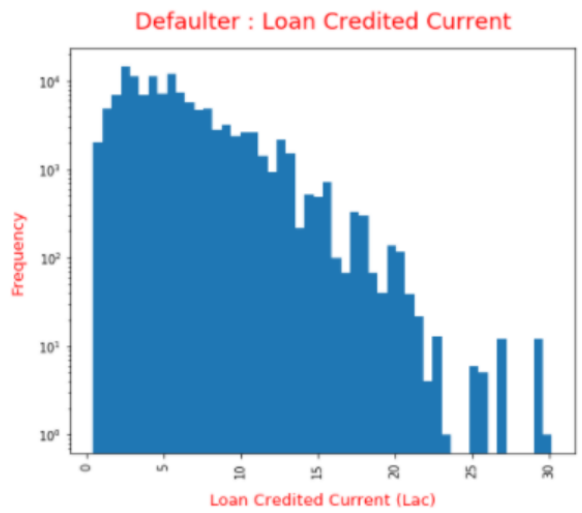
The insight and the result of the analysis of each variable are mentioned in upcoming slides.

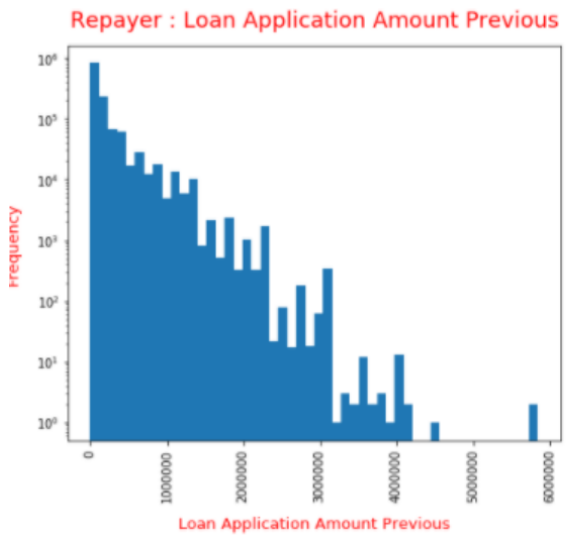
Univariate Analysis of merged previous and current application data for few variables:

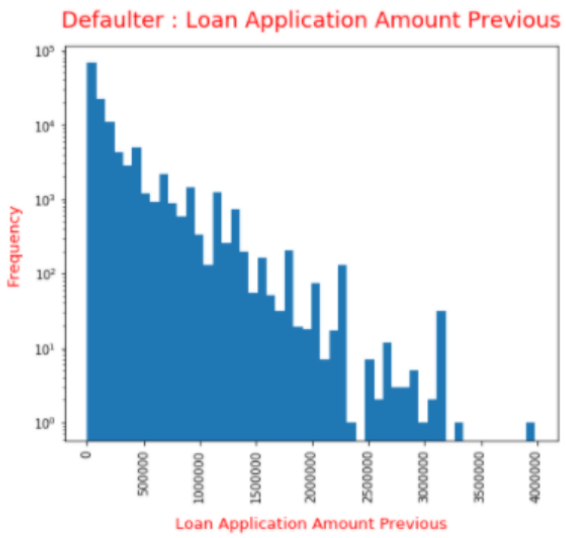


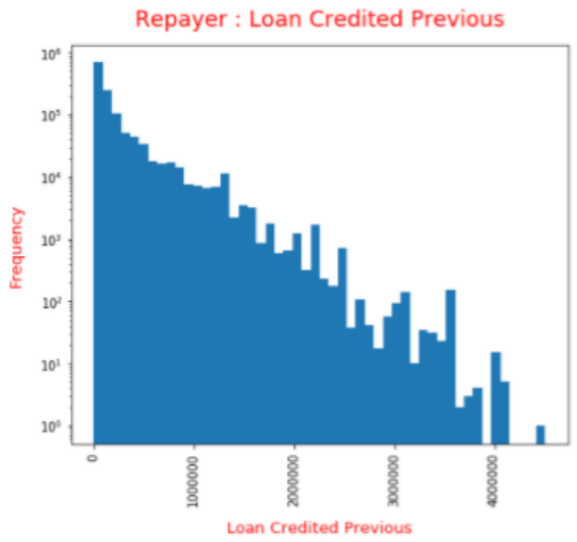


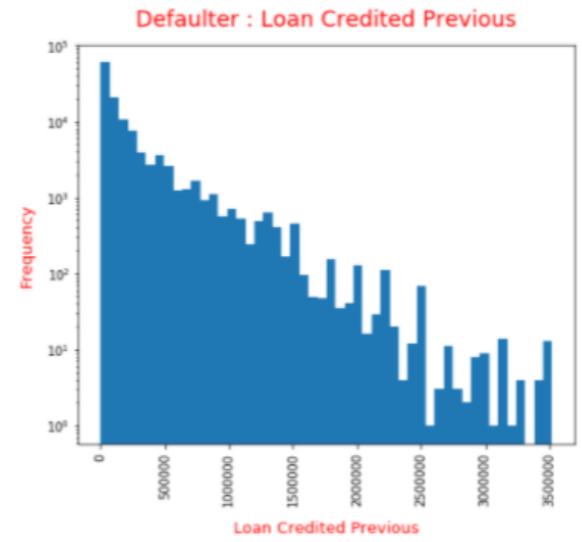




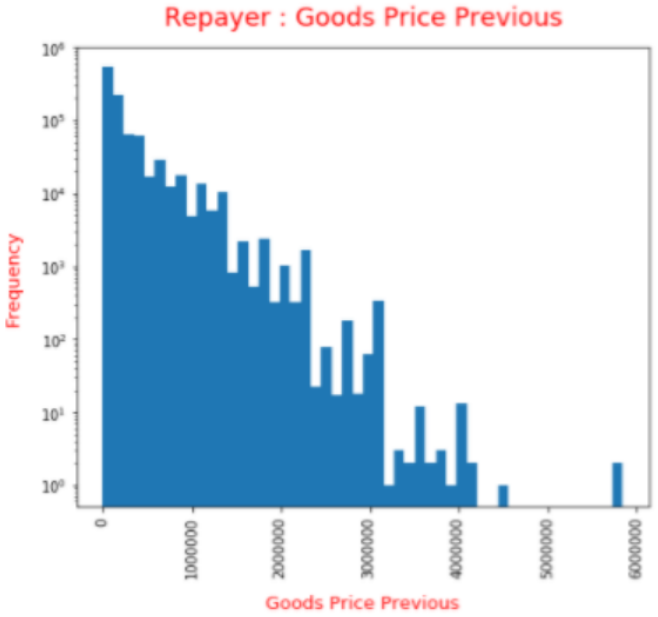


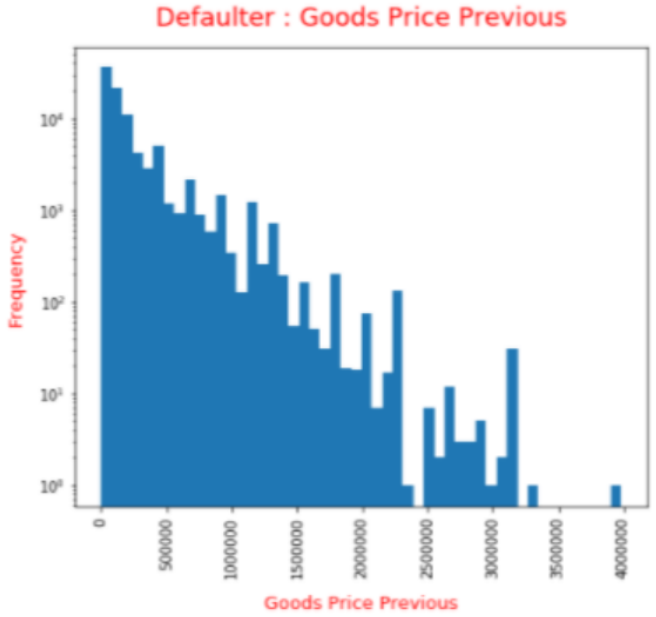


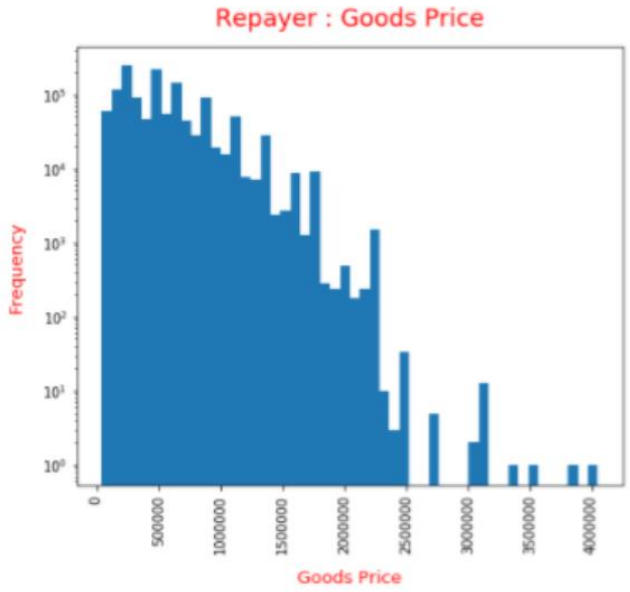


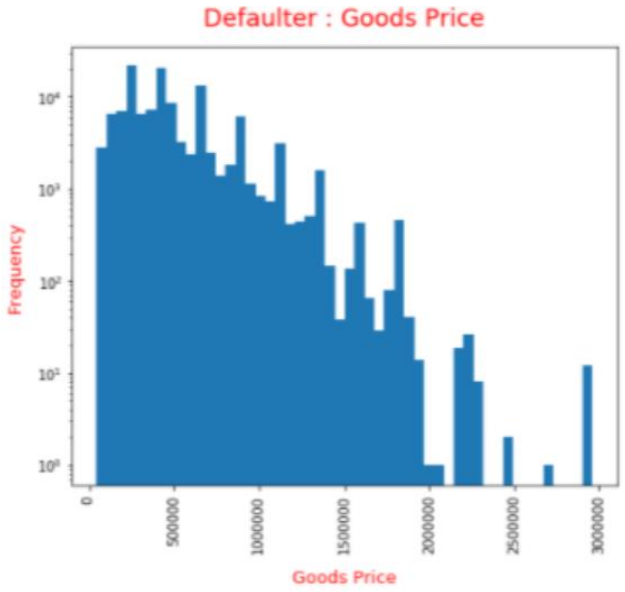


Univariate Analysis of merged previous and current application data for few variables: (Continue..)



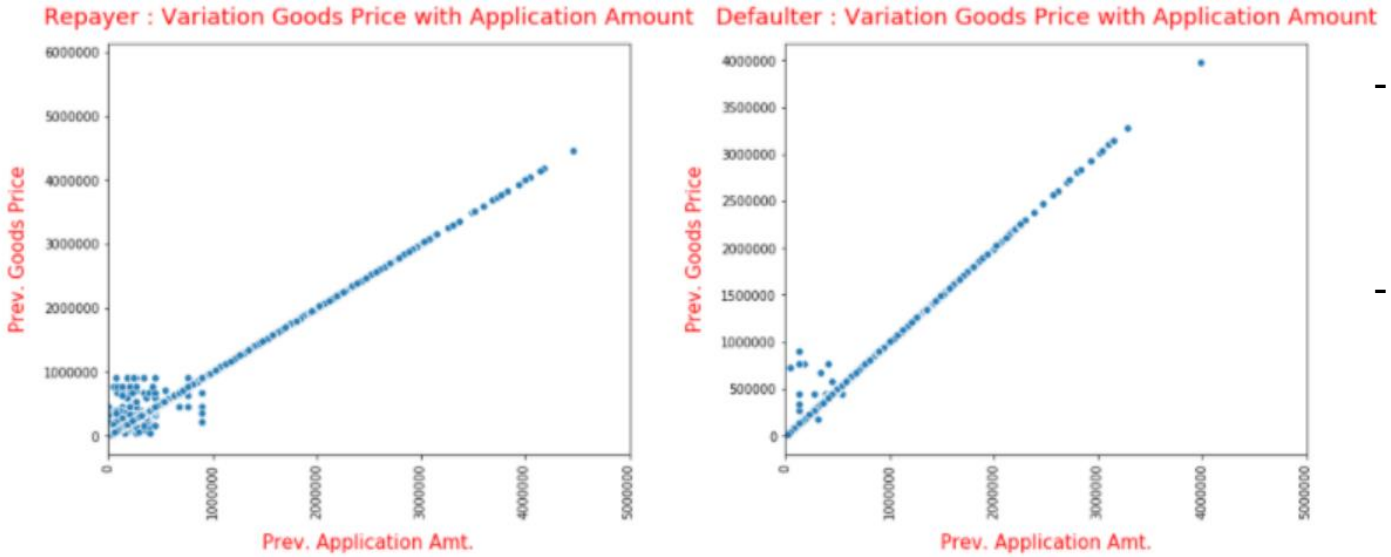






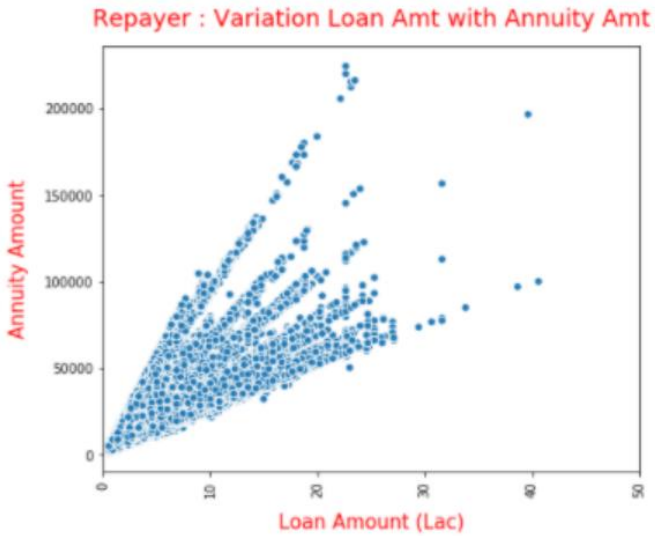
- Most of the cases the distribution for defaulter and re-payers are overlapping.
-
- Hence, we **cannot conclude any driving factors** from these frequency distributions.

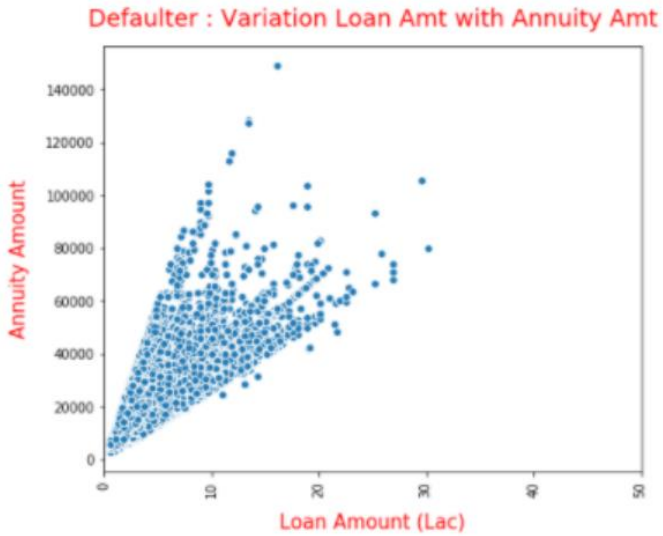
Bivariate Analysis of merged previous and current application data for few numerical variables :



- We can observed that customers in **both defaulter and re-payer category**, applied for **higher** loan amount for higher goods prices in their previous loan applications.
-

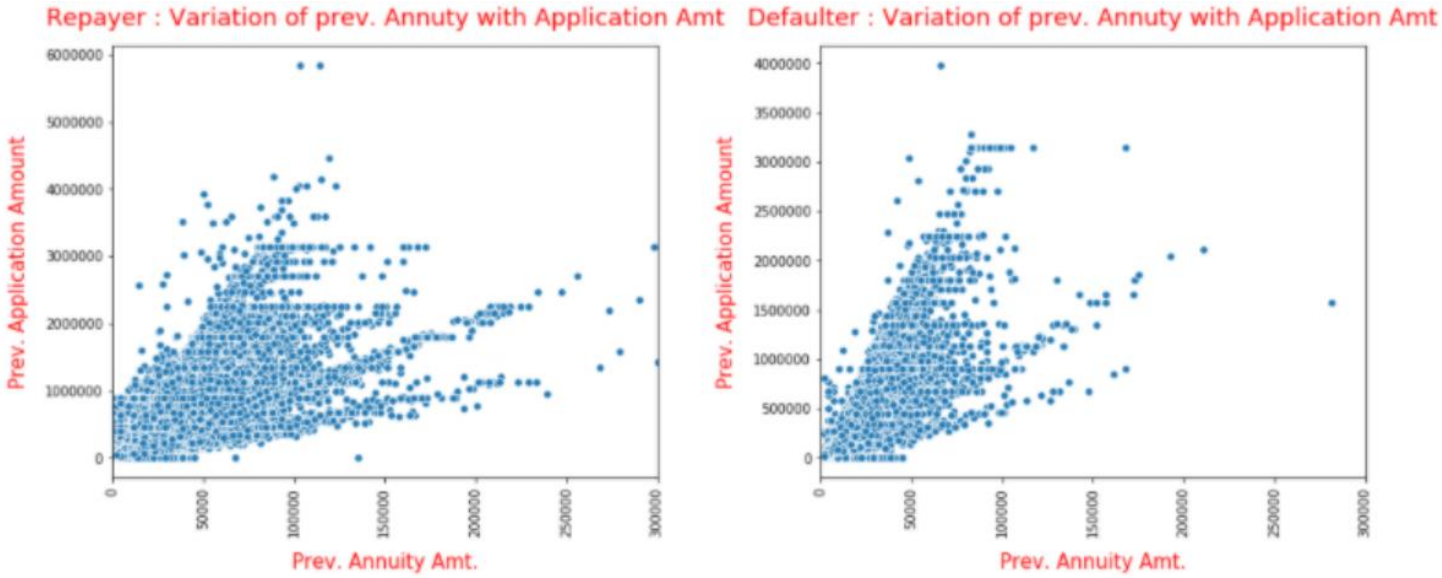
- **Customer who are most like repay the loan on time**, previously applied higher loan amount for costly goods than defaulters.





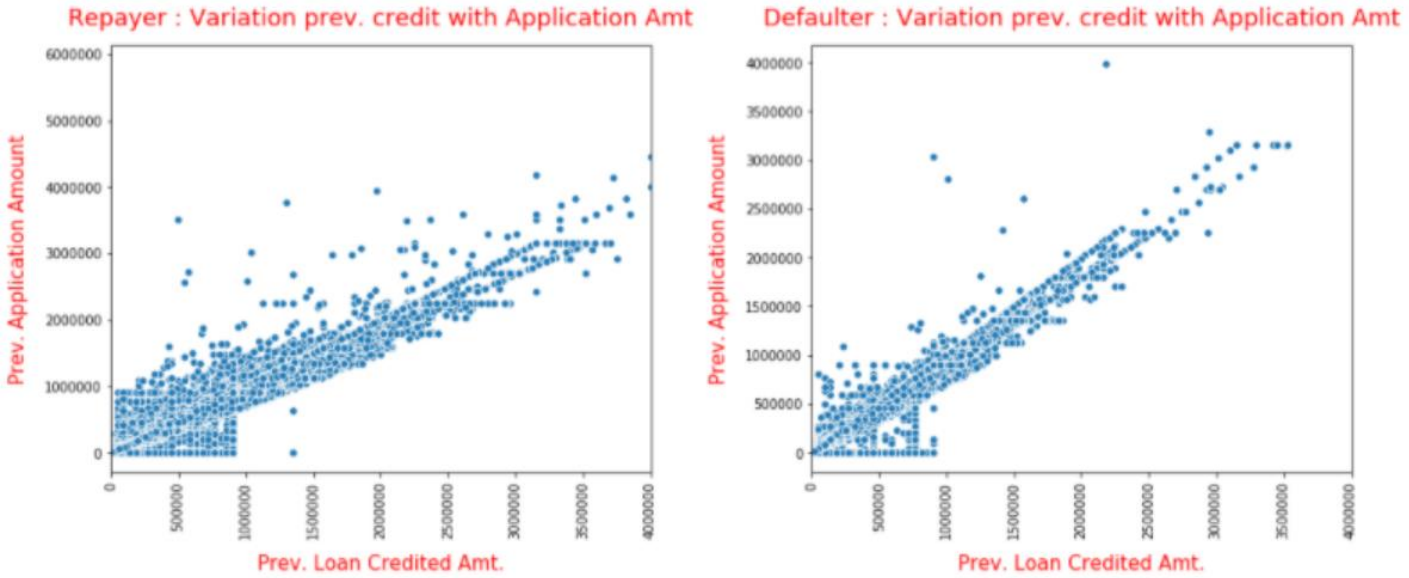
- We can observed that **customers with payment difficulty** pay lesser annuity amount(< 100K) with respect to re-payers with in loan amount range (0-30Lac.)`.

Bivariate Analysis of merged previous and current application data for few numerical variables : (Continue..)



- **Customers with payment difficulty, paid lesser annuity amount(< 100K) with respect to re-payers** in their previous application, while the loan application was for higher amount(1000k - 3000k)

- Re-payers paid **higher annuity (> 100k) for the loan application amount(1000k - 3000k)** in their previous application.

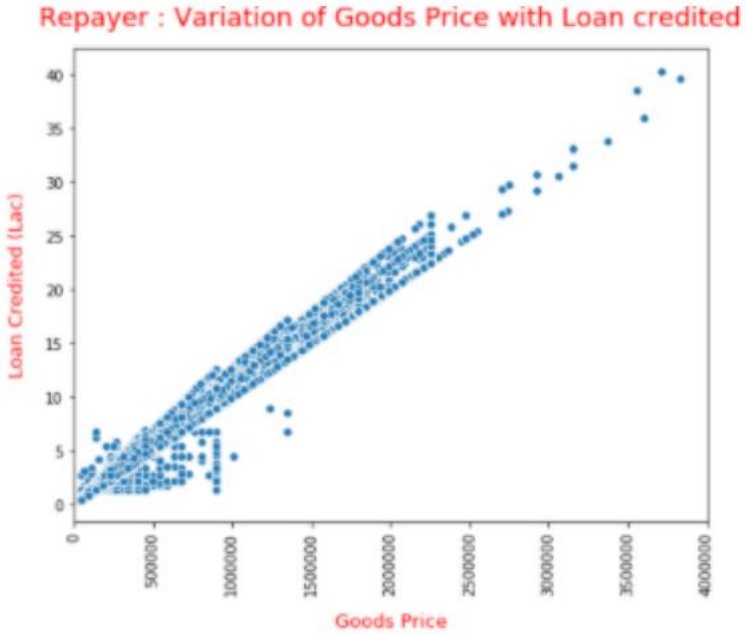


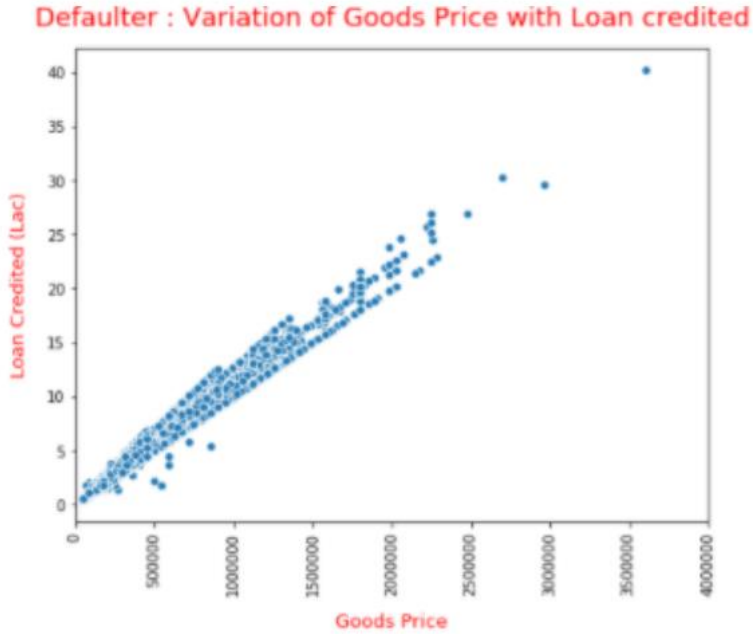
- **Customers who has not applied for any loan, got some loan created (0-100k) in previous loan application.**

- There are few **loan application between 2500k - 4000k for the defaulters**, compared to loan re-payers in previous loan data.

- There are **few loan application(1000k-4000k), for which re-payers got less credit in their previous application**`, compared to defaulter.

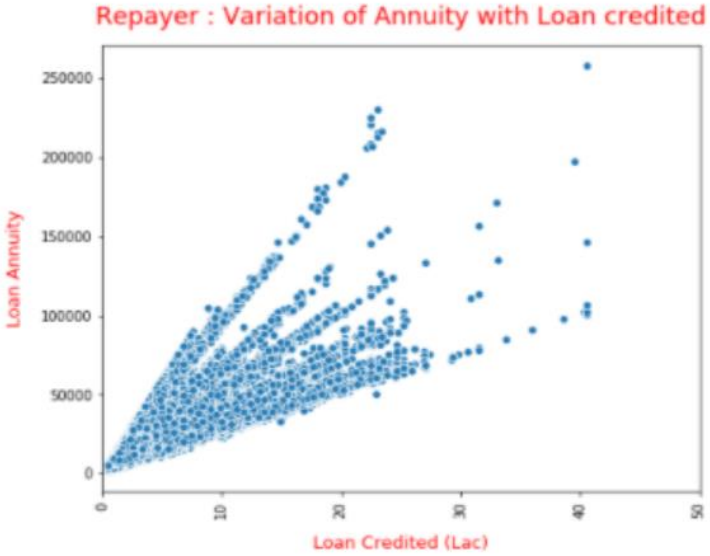
Bivariate Analysis of Current application data for few numerical variables :

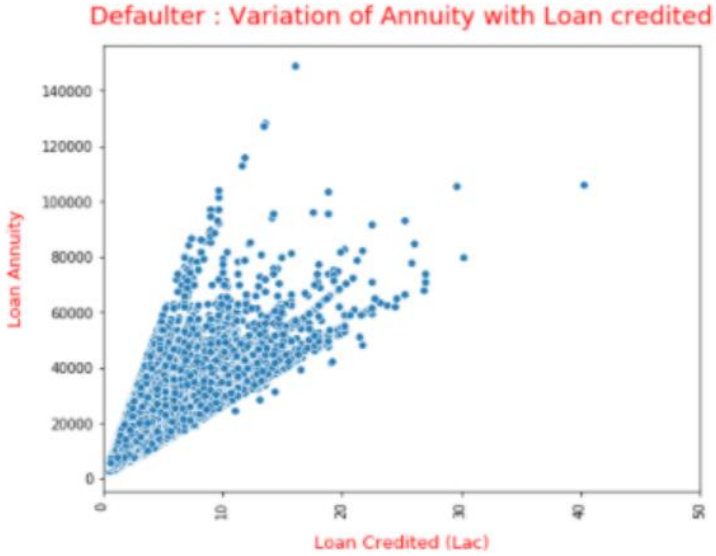




- Customer who bought **goods with higher price and have made payments on time , got higher loan credits than those with higher goods price but didn't pay loan**.

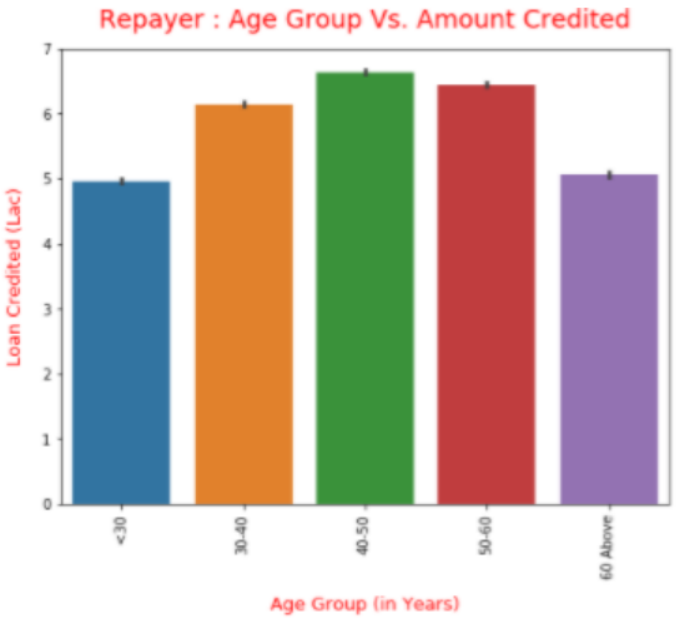
- **AMT_CREDIT** and **AMT_GOODS_PRICE** are highly correlated as based on the scatterplot where most of the data are consolidated in form of a line.

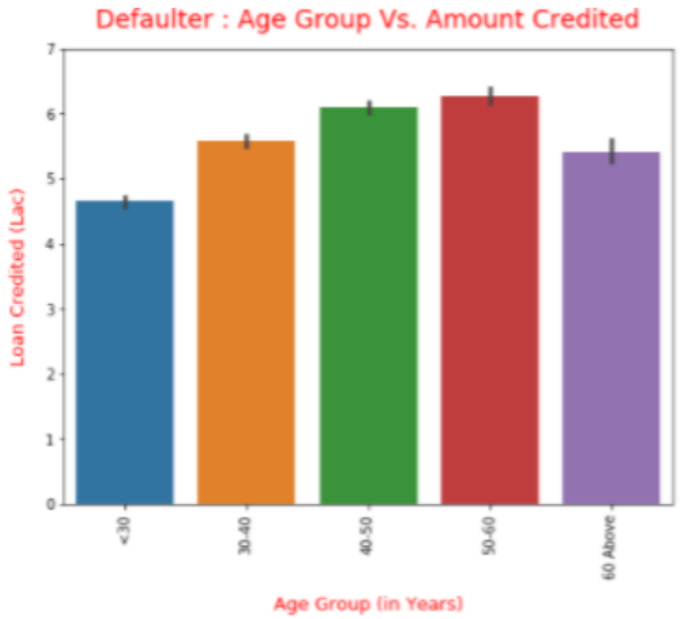




- **Customer who got higher loan credit (> 30Lac) has high Annuity (a fixed sum of money paid to bank each year) rate** and most likely to pay loan on time than the defaulters.

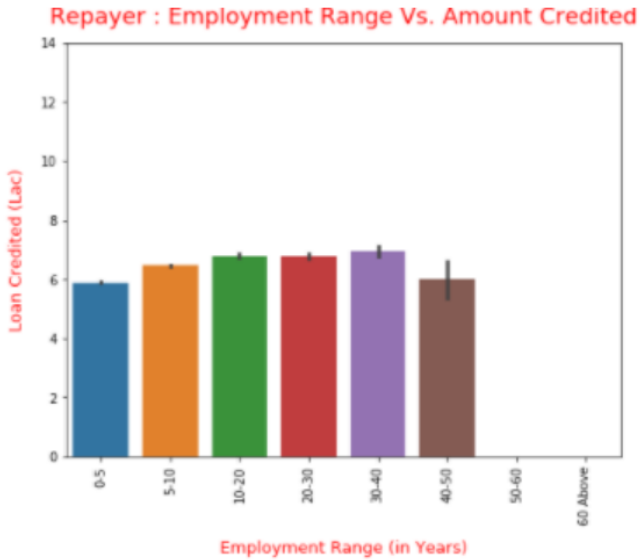
Bivariate Analysis of Current application data for few categorical variables :

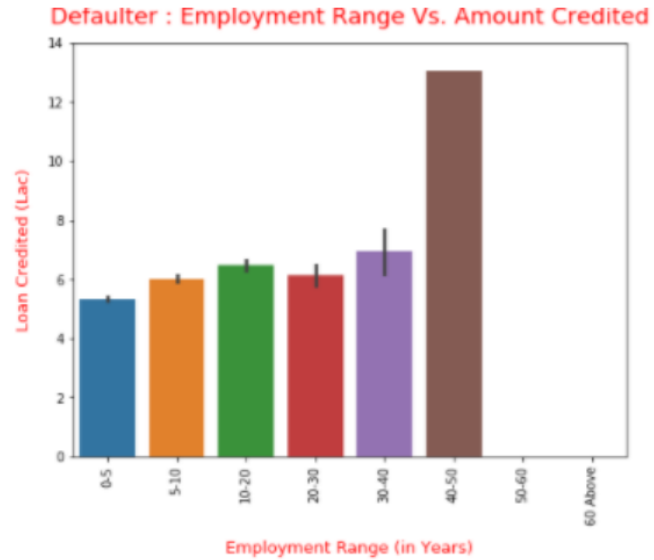




- Customer who are **able to repay the loan on time and credited higher loan amount, belonged to age group of 40-50 (years)**.

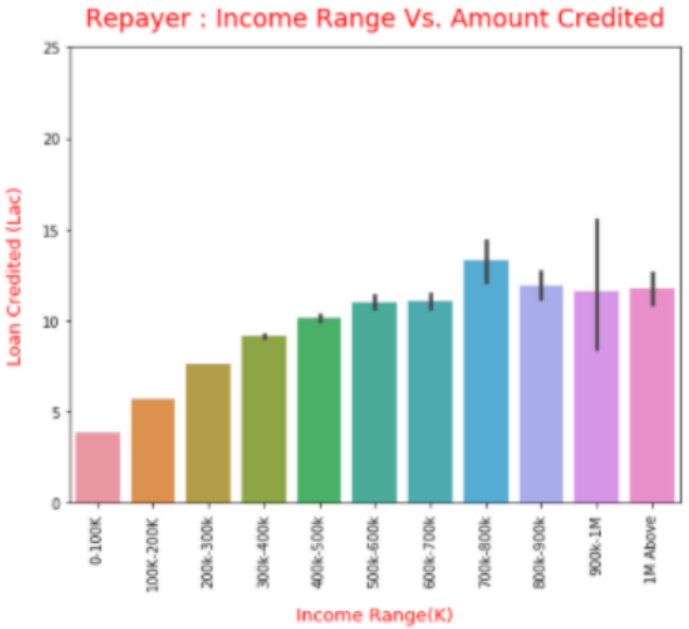
- Where as Customer who are **unable to repay the loan on time (defaulters) and credited higher loan amount, belonged to age group of 50-60 (years)**.

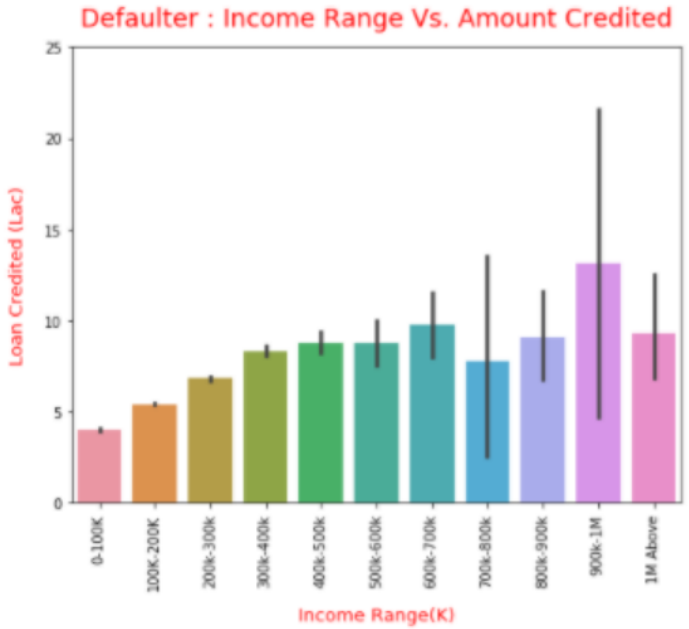




- Customer who are **able to repay the loan on time and credited higher loan amount, are having 30-40 years of employment tenure**.
- Where as Customer who are **unable to repay the loan on time (defaulters) and credited higher loan amount, are having 40-50 years of employment tenure**.

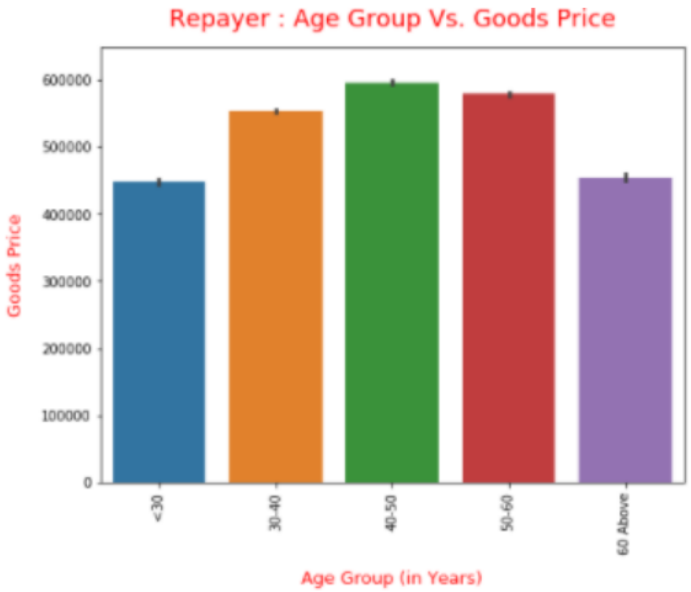
Bivariate Analysis of Current application data for few categorical variables : (Continue..)

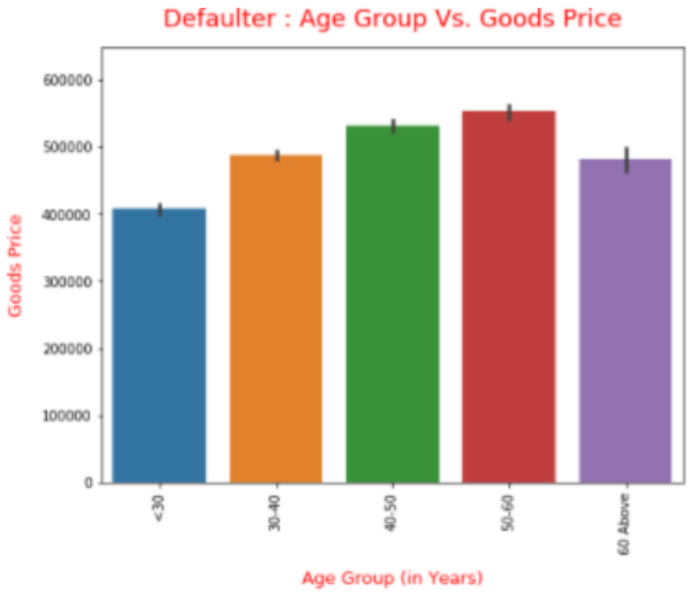




- Customer who are **able to repay the loan on time having income range between 700k-800k**, were credited higher loan amount.

- Where as Customer who are **unable to repay the loan on time (defaulters) and credited higher loan amount**, are having income range between **900k-1M**.





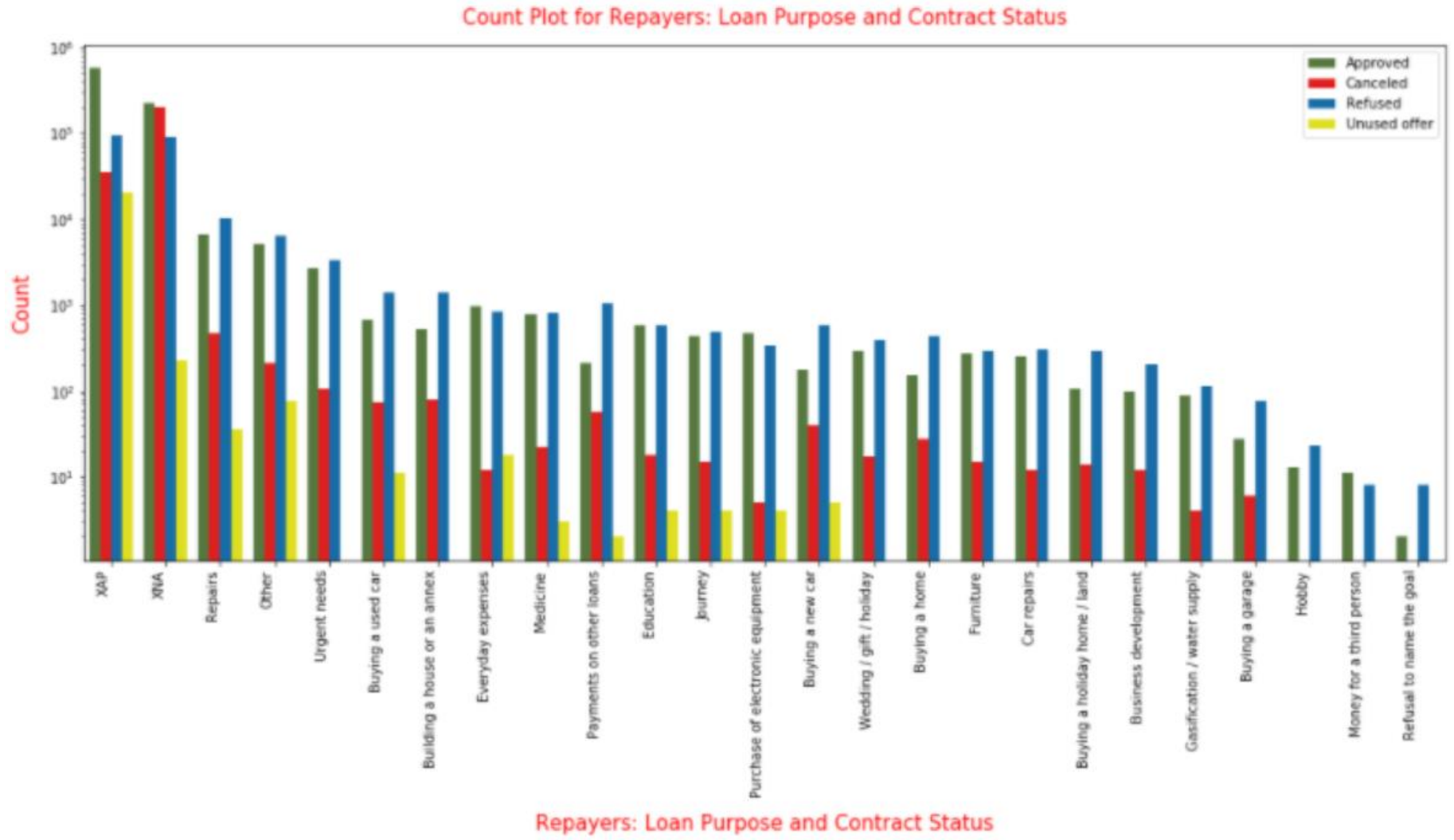
- Customer who are able to **repay loan on time are in age group between 40 -50 ,also bought goods with higher price**.

- Whereas **Customer who are defaulters belongs to age group between 40-50 ,bought goods with lower price** with respect to re-payers.

- Customer who are **defaulters belongs to age group between 50-60 ,bought goods with higher price**.

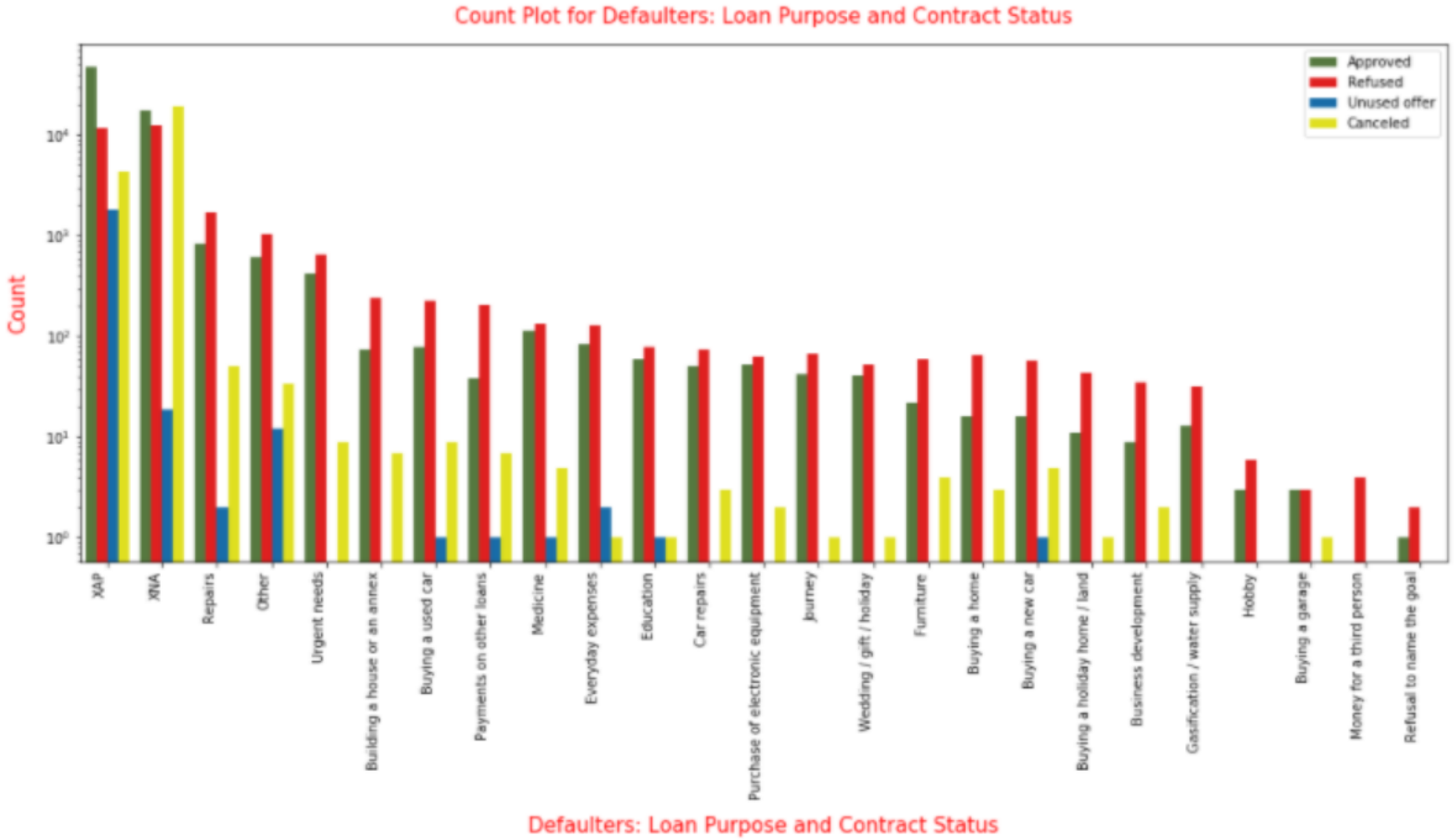
Bivariate Analysis of Merged dataset for categorical variables (TARGET = 0):

- NAME_CASH_LOAN_PURPOSE
- NAME_CONTRACT_STATUS



Bivariate Analysis of Merged dataset for categorical variables (TARGET = 1):

- NAME_CASH_LOAN_PURPOSE
- NAME_CONTRACT_STATUS



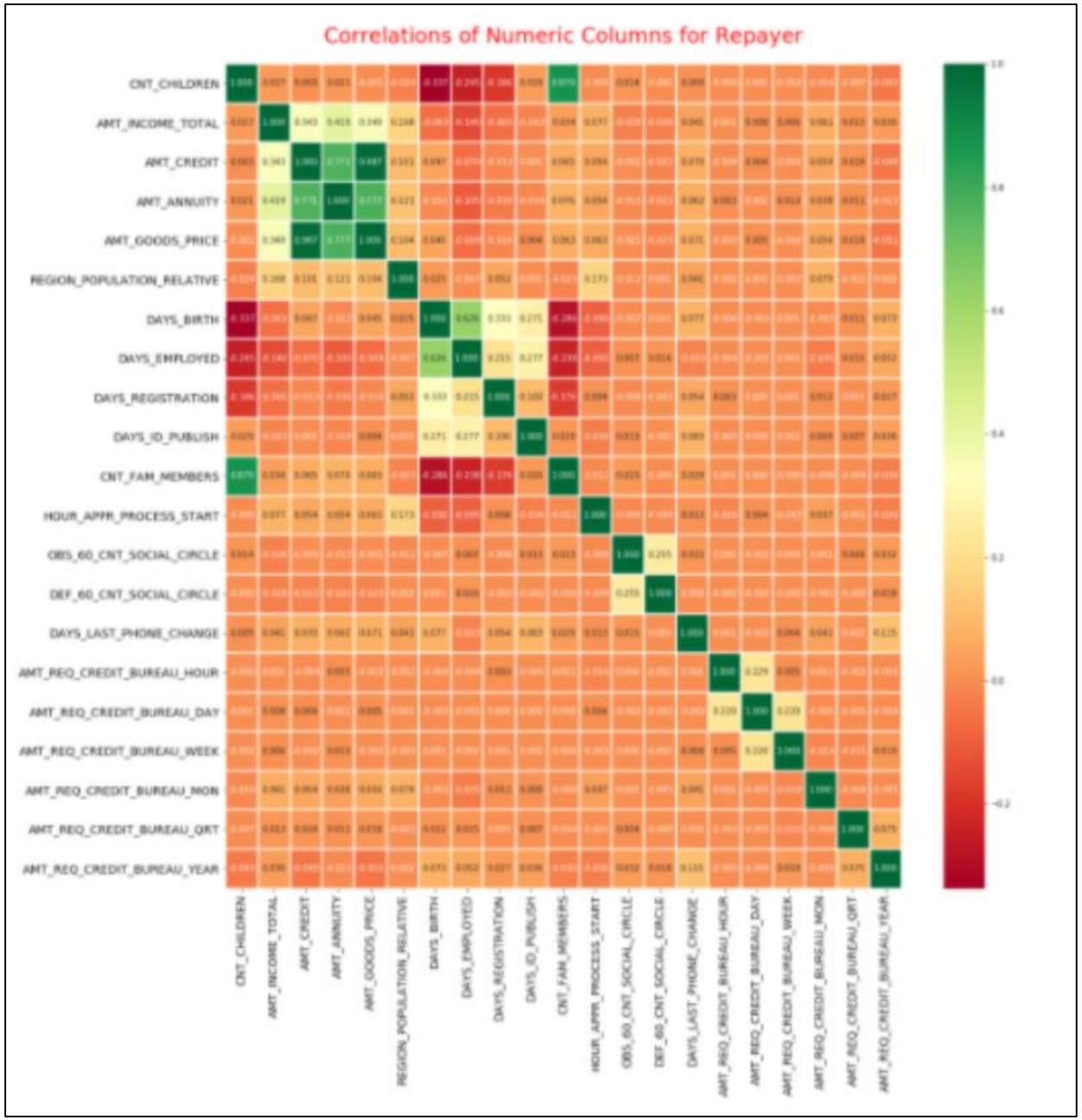
- **Loan purpose has high number of unknown values (XAP, XNA)**

- **Loan taken for the purpose of Repairs seems to have highest default rate**

- **A very high number application have been rejected by bank or refused by client** which has purpose as repair or other. This shows that purpose repair is taken as high risk by bank and either they are rejected or bank offers very high loan interest rate which is not feasible by the clients, thus they refuse the loan.

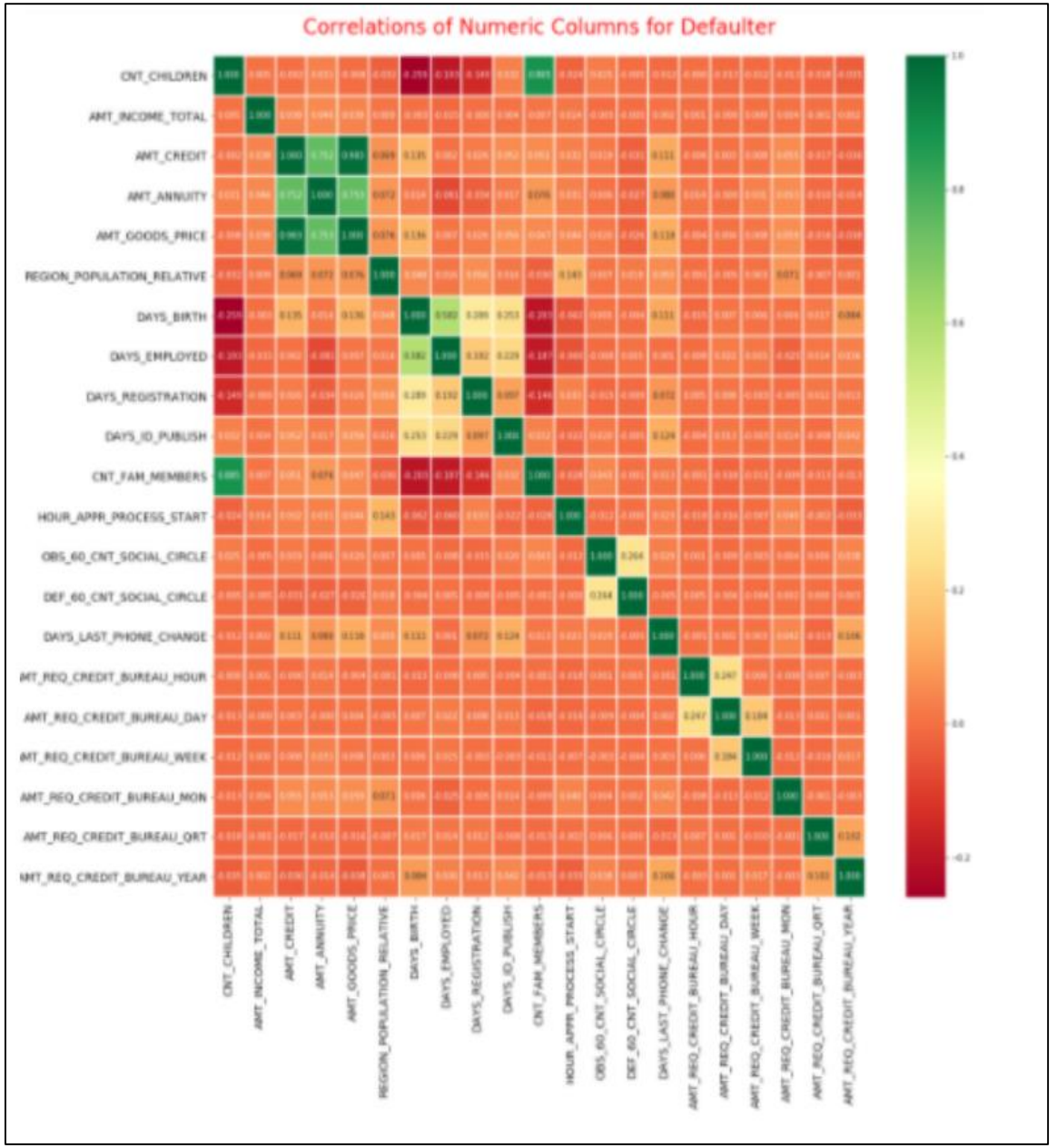
Top 10 correlation in current application data set for TARGET = 0

	Variable 1	Varibale 2	Correlation
86	AMT_GOODS_PRICE	AMT_CREDIT	0.987250
210	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
87	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
65	AMT_ANNUITY	AMT_CREDIT	0.771309
153	DAYS_EMPLOYED	DAYS_BIRTH	0.626114
64	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953
85	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349462
43	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799
126	DAYS_BIRTH	CNT_CHILDREN	0.336966
174	DAYS_REGISTRATION	DAYS_BIRTH	0.333151



Top 10 correlation in current application data set for TARGET = 1

	Variable 1	Varibale 2	Correlation
86	AMT_GOODS_PRICE	AMT_CREDIT	0.983103
210	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
87	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699
65	AMT_ANNUITY	AMT_CREDIT	0.752195
153	DAYS_EMPLOYED	DAYS_BIRTH	0.582185
174	DAYS_REGISTRATION	DAYS_BIRTH	0.289114
285	DEF_60_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.264159
126	DAYS_BIRTH	CNT_CHILDREN	0.259109
195	DAYS_ID_PUBLISH	DAYS_BIRTH	0.252863
351	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_HOUR	0.246741



Insight from top 10 correlation in current application data set for TARGET = 1 and TARGET = 0

TARGET = 1

- Credit amount is highly correlated with amount of goods price in case for defaulters.
  But the loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to re-payers(0.77).

- We can also see that re-payers have high correlation in number of days employed(0.62) when compared to defaulters(0.58).

- There is a severe drop in the correlation between total income of the client and the credit amount(0.038) amongst defaulters whereas it is 0.342 among re-payers.

- Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in re-payers.

- There is a slight increase in defaulted to observed count in `social circle among defaulters(0.264) when compared to re-payers(0.255).

TARGET = 0

- Credit amount is highly correlated with Amount of goods price , Loan annuity and with total income.

- Loan annuity is highly correlated with Amount of goods price, Credit amount and with total income.

- Number of days employed highly correlated for re-payer.

Decisive Factors whether an applicant will be Repayer:

- `NAME_EDUCATION_TYPE:` Academic degree and Higher education Degree holders has less defaults.
- `NAME_INCOME_TYPE:` Student ,Businessmen and Pensioner have no defaults
- `DAYS_BIRTH:` Customer above age of 50 have low probability of defaulting
- `DAYS_EMPLOYED:` Clients with 40+ year experience having less than 1% default rate
- `AMT_INCOME_TOTAL:` Applicant with Income more than 700,000 are less likely to default
- `NAME_CONTRACT_TYPE:` Applicant with contract type of Revolving Fund are less likely to default.
- `NAME_FAMILY_STATUS:` Applicant with family status as Widow less likely to default.
- `OCCUPATION_TYPE:` Applicant with occupation type Manager are less likely to default.As The Laborers , Sales Staff, Core Staff, Managers and Drivers are the highest number ofapplicants. TheLaborers turn out to be even higher defaulters aswell, Sales Staff is also higher in count of defaulters. Managers on the other hand are slightly low in count in being defaulter thanrepayers. According to this information , Managers can be preferred higher over laborers and Sales Staff.
- `NAME_PORTFOLIO:` In Previous and current application merged data POS seems to be better in terms of repayer.
- `NAME_PRODUCT_TYPE:` XNA (unknow product type) and x-sell which are more of repay.
- `NAME_SELLER_INDUSTRY:` Consumer electronics is better at repaying.

- `Top 10 Correlation` in application dataframe for repayers are below:
  - `AMT_GOODS_PRICE and AMT_CREDIT`
  - `CNT_FAM_MEMBERS and CNT_CHILDREN`
  - `AMT_GOODS_PRICE and AMT_ANNUITY`
  - `AMT_ANNUITY and AMT_CREDIT`
  - `1DAYS_EMPLOYED and DAYS_BIRTH`
  - `AMT_ANNUITY and AMT_INCOME_TOTAL`
  - `AMT_GOODS_PRICE and AMT_INCOME_TOTAL`
  - `AMT_CREDIT and AMT_INCOME_TOTAL`
  - `DAYS_BIRTH and CNT_CHILDREN`
  - `DAYS_REGISTRATION and DAYS_BIRTH`

Decisive Factors whether an applicant will be Defaulter:

- `CODE_GENDER:` Men are at relatively higher default rate
- `NAME_FAMILY_STATUS :` People who have civil marriage or who are single default a lot.
- `NAME_EDUCATION_TYPE:` People with Secondary education default a lot.
- `OCCUPATION_TYPE:` Avoid Low-skill Laborers, Drivers , Laborers and Cooking staff as the default rate is huge.
- `DAYS_BIRTH:` Avoid young people who are in age group of < 30-40 as they have higher probability of defaulting.
- `DAYS_EMPLOYED:` People who have 0-5 years of employment have high default rate.
- `NAME_PORTFOLIO:` In Previous and current application merged data XNA and CARDS are more towards defaulters.
- `NAME_PRODUCT_TYPE:` Walk-in type is observed to be high at turning into defaulter.
- `NAME_SELLER_INDUSTRY:` While XNA(unknow induxtry type) is amajor defaulter turning category.
- `FLAG_OWN_CAR` There are more applicants who do not have car who have payment difficulties.
- `NAME_TYPE_SUITE:` Unaccompanied show a higher Defaulter rate.
- `NAME_TYPE_SUITE_PREV:` In the previous data,the same pattern is seen as applclication data. Unaccompaniedand more of defaulters.
- `Top 10 Correlation` in application dataframe for Defaulters are below:
  - `AMT_GOODS_PRICE and AMT_CREDIT`
  - `CNT_FAM_MEMBERS and CNT_CHILDREN`
  - `AMT_GOODS_PRICE and AMT_ANNUITY`
  - `AMT_ANNUITY and AMT_CREDIT`
  - `DAYS_EMPLOYED and DAYS_BIRTH`
  - `DAYS_REGISTRATION and DAYS_BIRTH`
  - `DEF_60_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE`
  - `DAYS_BIRTH and CNT_CHILDREN`
  - `DAYS_ID_PUBLISH and DAYS_BIRTH`
  - `AMT_REQ_CREDIT_BUREAU_DAY and AMT_REQ_CREDIT_BUREAU_HOUR`

The following attributes indicate that customers from these category tend to default but then due to the more number of customers and the amount of loan, the bank could provide loan with higher interest to mitigate any default risk thus preventing business loss:

- `NAME_HOUSING_TYPE:` High number of loan applications are from the category of people who live in Rented apartments & living with parents and hence offering the loan with higher interest rate would mitigate the loss if any of those defaults.
- `AMT_CREDIT:` People who get loan for 300-600k tend to default more than others and hence having higher interest specifically for this credit range would be ideal.
- `AMT_INCOME:` Since most percentage of the applicants have Income total less than 300K and they have high probability of defaulting, they could be offered loan with higher interest compared to other income category.
- `NAME_CASH_LOAN_PURPOSE:` Loan taken for the purpose of Repairs seems to have highest default rate. A very high number applications have been rejected by bank or refused by client in previous applications as well which has purpose as repair. This shows that purpose repair is taken as high risk by bank and either they are rejected, or bank offers very high loan interest rate which is not feasible by the clients, thus they refuse the loan. The same approach could be followed in future as well.

-END-