

# Credit EDA Case Study

## Data Quality checks and handling missing values:

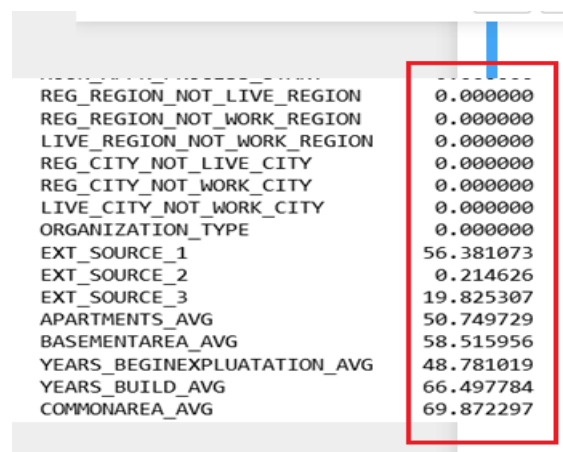
Two data sets were provided as part of this case study

- Application Data
- Previous application data

In order to perform data quality checks and handling missing values, **we have considered Application data set and performed necessary actions as explained below.**

## To find out the missing values and handle it :

- We have observed there were many missing values in this data set, so using the indexing and count of missing values in each column, identified the % of missing values for each column.



REG_REGION_NOT_LIVE_REGION	0.000000
REG_REGION_NOT_WORK_REGION	0.000000
LIVE_REGION_NOT_WORK_REGION	0.000000
REG_CITY_NOT_LIVE_CITY	0.000000
REG_CITY_NOT_WORK_CITY	0.000000
LIVE_CITY_NOT_WORK_CITY	0.000000
ORGANIZATION_TYPE	0.000000
EXT_SOURCE_1	56.381073
EXT_SOURCE_2	0.214626
EXT_SOURCE_3	19.825307
APARTMENTS_AVG	50.749729
BASEMENTAREA_AVG	58.515956
YEARS_BEGINEXPLUATATION_AVG	48.781019
YEARS_BUILD_AVG	66.497784
COMMONAREA_AVG	69.872297

- And dropped all columns from data frame for which missing values % is more than 50. We have also found few more columns which are having around 47% missing values. Since these are almost around 50% ,we have removed these columns as well.
- To extend the data enhancements and maintain a data frame with accurate values, we have identified columns with NULL values and even applied the same 47% logic and removed those columns from the data frame.

After that there are still some columns left with some null values, we will impute the ones where null percentage is less than 1 %

### We have imputed the below columns with mean and mode

- AMT\_GOODS\_PRICE (integer)
- NAME\_TYPE\_SUITE(Object)
- EXT\_SOURCE\_2(integer)

As shown in the below diagram, finding the mean and median for both **AMT\_GOODS\_PRICE** and **EXT\_SOURCE\_2** and based on percentile variation between 25<sup>th</sup> and 75<sup>th</sup>, let's impute the missing values by mean values of **AMT\_GOODS\_PRICE** and **EXT\_SOURCE\_2** respectively.

```
#checking the statistics for below columns
df[['AMT_GOODS_PRICE', 'EXT_SOURCE_2']].describe().T
```

	count	mean	std	min	25%	50%	75%	max
AMT_GOODS_PRICE	307233.0	538396.207429	369446.46054	4.050000e+04	238500.000000	450000.000000	679500.000000	4050000.000
EXT_SOURCE_2	306851.0	0.514393	0.19106	8.173617e-08	0.392457	0.565961	0.663617	0.855

The difference between the mean and the median (lower the better), and the variation from 25th to 75th percentile (quite small in this case).  
Thus, let's impute the missing values by the mean value of AMT\_GOODS\_PRICE and EXT\_SOURCE\_2 respectively.

Performing the similar operation even for **NAME\_TYPE\_SUITE** column as well. Using the mode value, replace the remaining NULL values as shown in the below diagram.

```
#using mode to replace null values
```

```
df['NAME_TYPE_SUITE'].fillna(df['NAME_TYPE_SUITE'].mode()[0], inplace=True)
df['NAME_TYPE_SUITE'].value_counts()
```

```
Unaccompanied    249818
Family            40149
Spouse, partner   11370
Children          3267
Other_B           1770
Other_A           866
Group of people   271
Name: NAME_TYPE_SUITE, dtype: int64
```

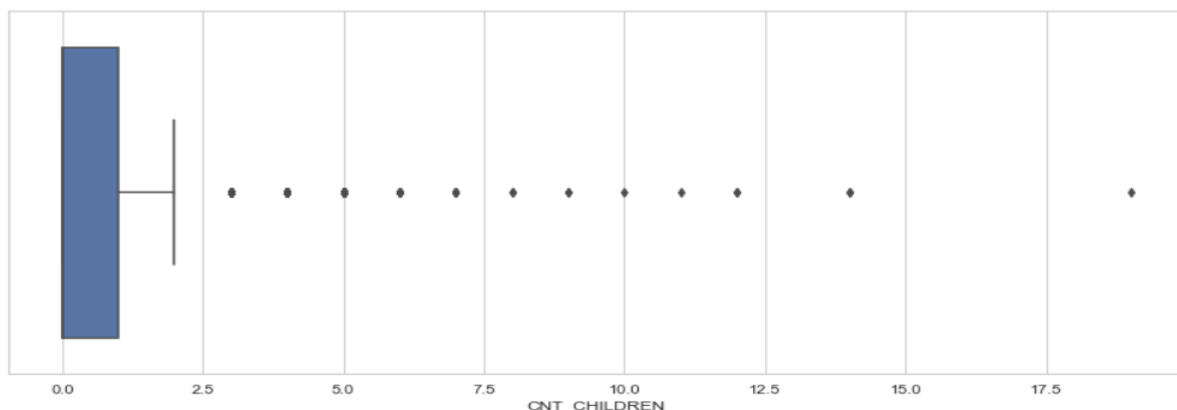
## Checking the Data frame for outlier values and analyze them

- Checking the outliers for columns and understanding the reason to mention that as an outlier.
- Here in our analysis to find out the outliers, we have considered few numerical columns and analyzed the statistics of them.
- **If we observe the below screenshot, there are 3 columns with outlier values** which are having a huge difference compared to the regular intervals of other values.

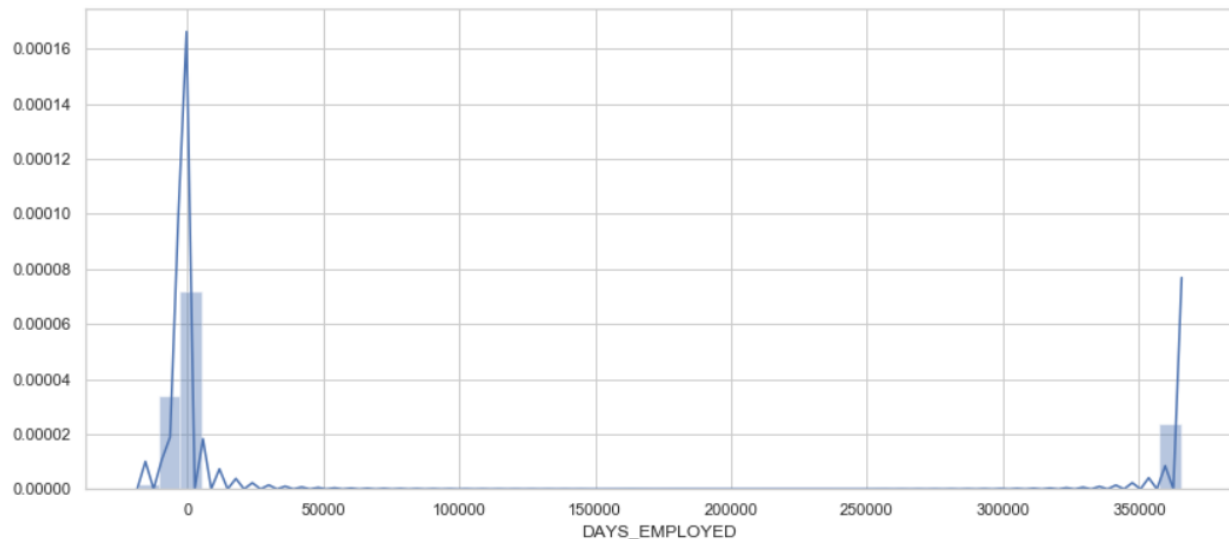
	count	mean	std	min	25%	50%	75%	max
AMT_INCOME_TOTAL	168738.0	179096.702061	303551.444817	26550.0	112500.00	157500.0	225000.0	117000000.0
AMT_CREDIT	168738.0	620729.363955	408611.456605	45000.0	284400.00	528633.0	835605.0	4050000.0
AMT_ANNUITY	168738.0	27891.026402	14464.318010	1980.0	17217.00	26014.5	35685.0	258025.5
CNT_CHILDREN	168738.0	0.512647	0.769343	0.0	0.00	0.0	1.0	19.0
AMT_GOODS_PRICE	168738.0	557586.695939	374748.321567	40500.0	247500.00	454500.0	702000.0	4050000.0
DAYS_BIRTH	168738.0	-14876.485095	3594.864088	-25200.0	-17601.75	-14688.0	-11969.0	-7676.0
DAYS_ID_PUBLISH	168738.0	-2871.611018	1500.781393	-7197.0	-4216.00	-2990.0	-1594.0	0.0
DAYS_EMPLOYED	168738.0	-2469.153759	2553.921340	-17912.0	-3294.00	-1719.0	-806.0	365243.0
DAYS_REGISTRATION	168738.0	-4636.211802	3247.769804	-22928.0	-6954.00	-4272.0	-1837.0	0.0

## Explanation using box plots for outlier values:

1. **CNT\_CHILDREN** : Here as shown in the below screenshot, there is one value 19 as per humans cannot have so many children hence we will also consider this is an outlier:



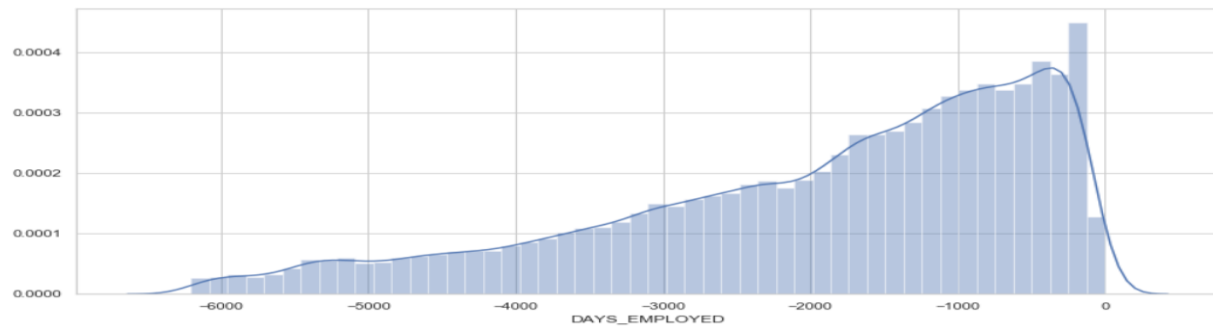
2. **DAYS\_EMPLOYED:** there is one value which is present here 365243. It might be due to manual error while data entry plus all the values in this column are negative except for this value, hence we will treat this as an outlier.



3. **AMT\_INCOME\_TOTAL:** As max amount is way above the mean and 75th percentile hence i will consider there are outliers present in this column.

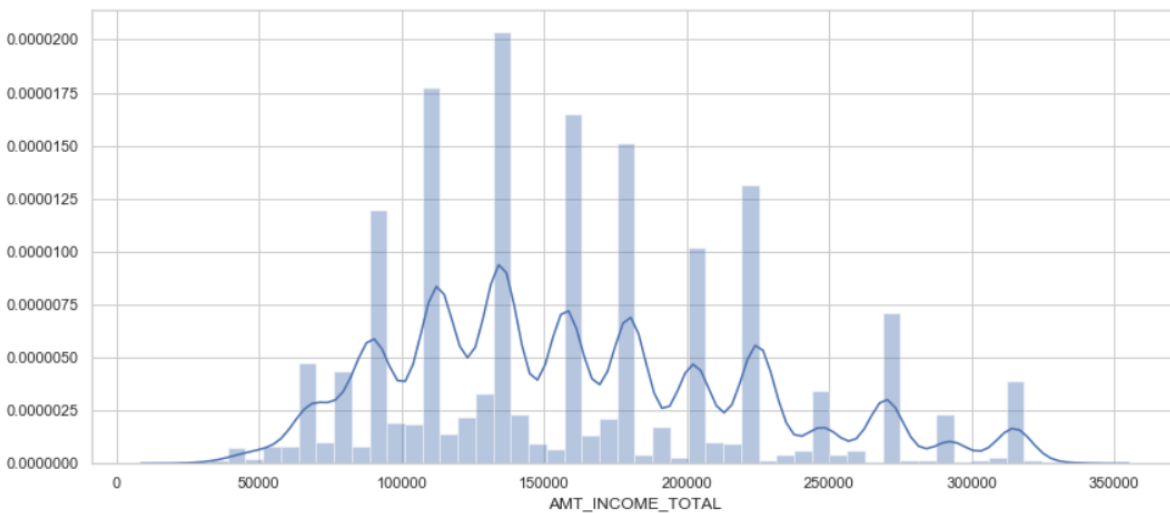
- As per the above plots, we have observed the outliers in 3 columns and now we will be removing those outliers and plot the same columns again to find the difference.
- We have defined a function to handle and remove outliers from specific columns using which we have removed outliers, this function is available in our python file from jupyter notebook.

**DAYS\_EMPLOYED** columns plotting after removing outliers in it.



Maximum Applicants lie in te early years of Employment while they have applied for loan.

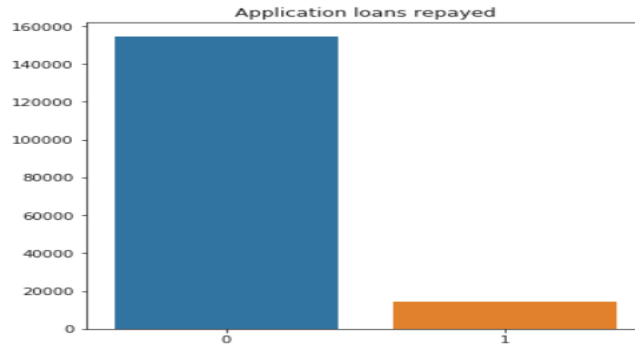
**AMT\_INCOME\_TOTAL** column after removing the outlier values in it.



Income ranges from 25k to 300k. Tere are few spikes in between ,this is the plot we get after removing outliers

## Data Analysis

- While working with the data frame based on Target variable, **As shown in the below screenshot, we can clearly see the imbalance between target type 1 and 0.**
- **Ratio is of 91.5 : 8.45**



- Devide the dataset into two subsets based on Target variable. i.e. Target=0 and Target=1.

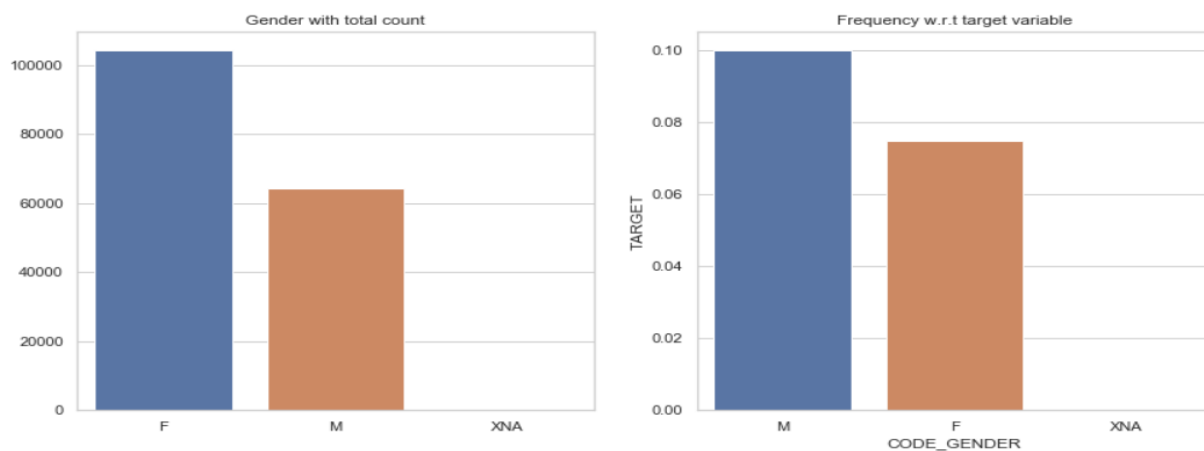
```
#segregating dataframe w.r.t Target variable
not_fraud=df_nonull[df_nonull.TARGET==0]
fraud=df_nonull[df_nonull.TARGET==1]
```

- Perform univariate analysis for categorical variables for both 0 and 1.

In order to perform Univariate analysis on categorical variables available in the data set for both 0 and 1, we have choosen few columns in terms of variable and compared against total number of rows I data frame and based on Target variable.

### CODE\_GENDER

As shown in the below screenshot when we plot the **Male and Female gender ratio for total number of rows and against Target variable we found different ratios.**



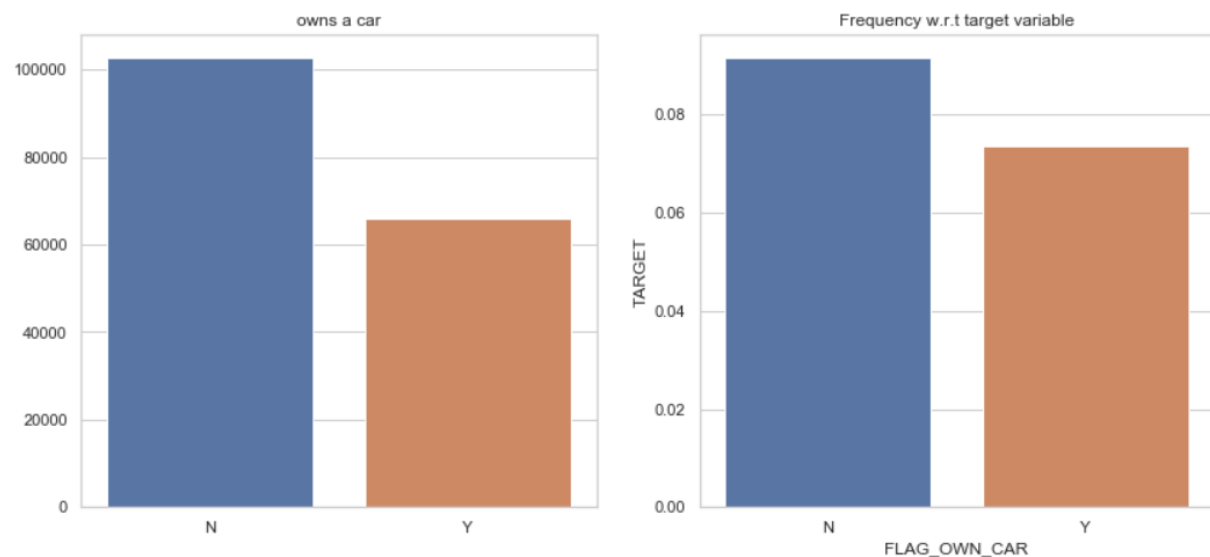
As shown in the above screenshot, we have plotted **CODE\_GENDER** values against total number of rows and Target variable.

We found that the number of female clients is almost double the number of male clients. Looking to the percent of defaulted credits, males have a higher chance of not returning their loans (~10%), comparing with women (~7%).

Contract type Revolving loans are just a small fraction from the total number of loans but at the same time, a larger amount of Revolving when compared with their frequency are not repaid.

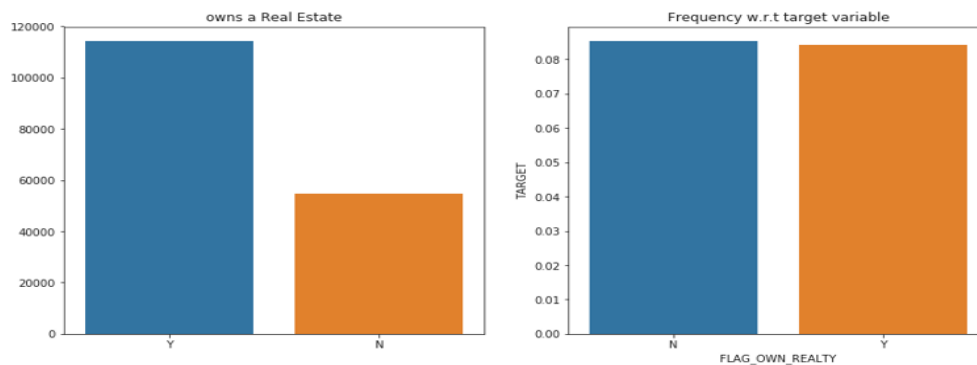
**We have performed the similar analysis on few other columns.**

### FLAG\_OWN\_CAR & Real Estate



The clients that owns a car are almost a half of the ones that doesn't own one. The clients that owns a car are less likely to not repay a car that the ones that own. **Both categories have not-repayment rates around 8%.**

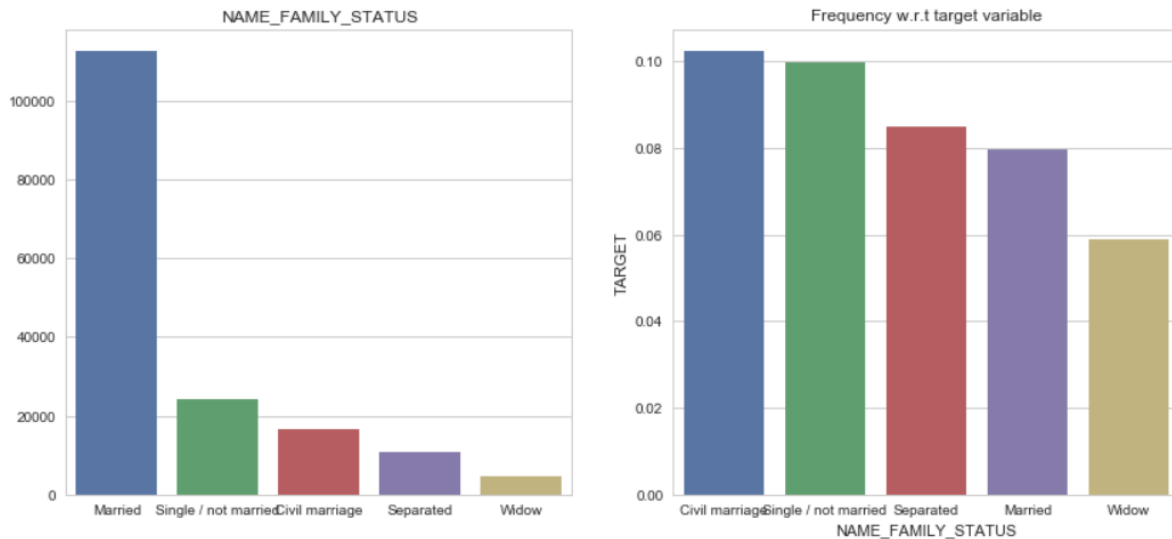
The clients that owns real estate are more than double of the ones that doesn't own. Both categories (owning real estate or not owning) have non-repayment rates near to 8%.



## NAME\_FAMILY\_STATUS

Most of clients are married, followed by Single/not married and civil marriage.

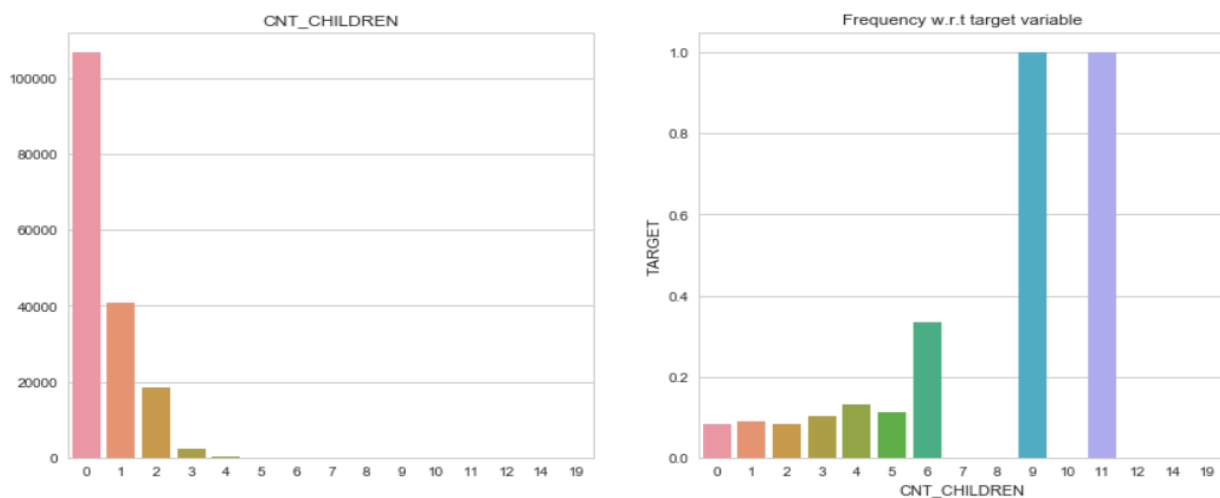
In terms of percentage of not repayment of loan, Civil marriage has the highest percent of not repayment with Widow the lowest.



## CNT\_CHILDREN

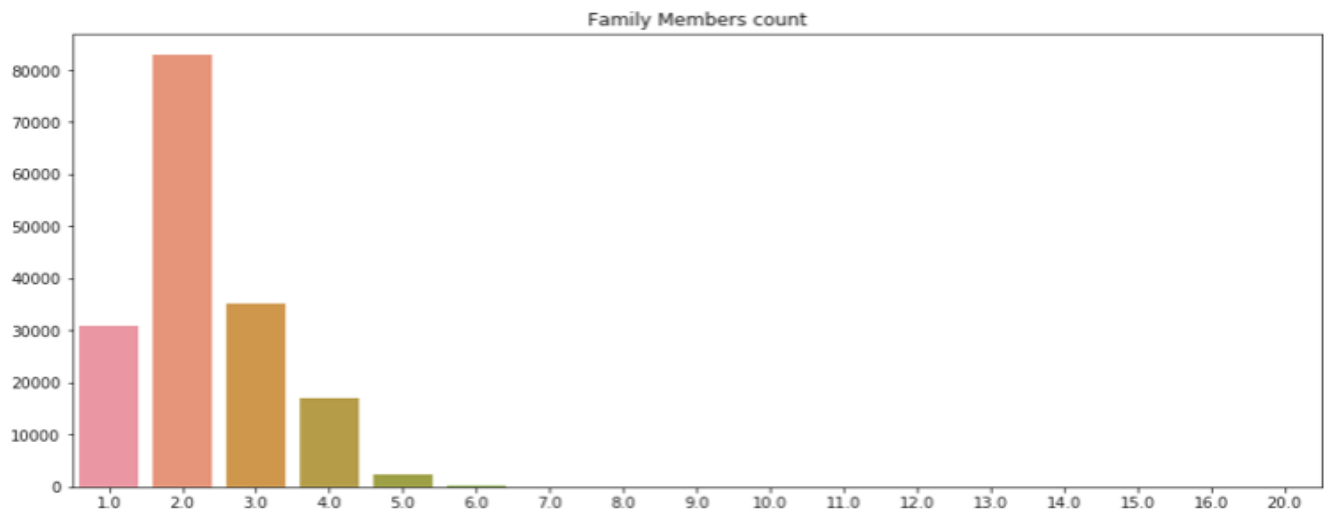
When we observed CNT\_CHILDREN column in our analysis, we observed that person with more than number of children then chances of loan payment become less. May be because of the income and number of dependents ratio in the home.

Also, A large majority of applicants did not had children when they applied for loan





## CNT\_FAM\_MEMBERS

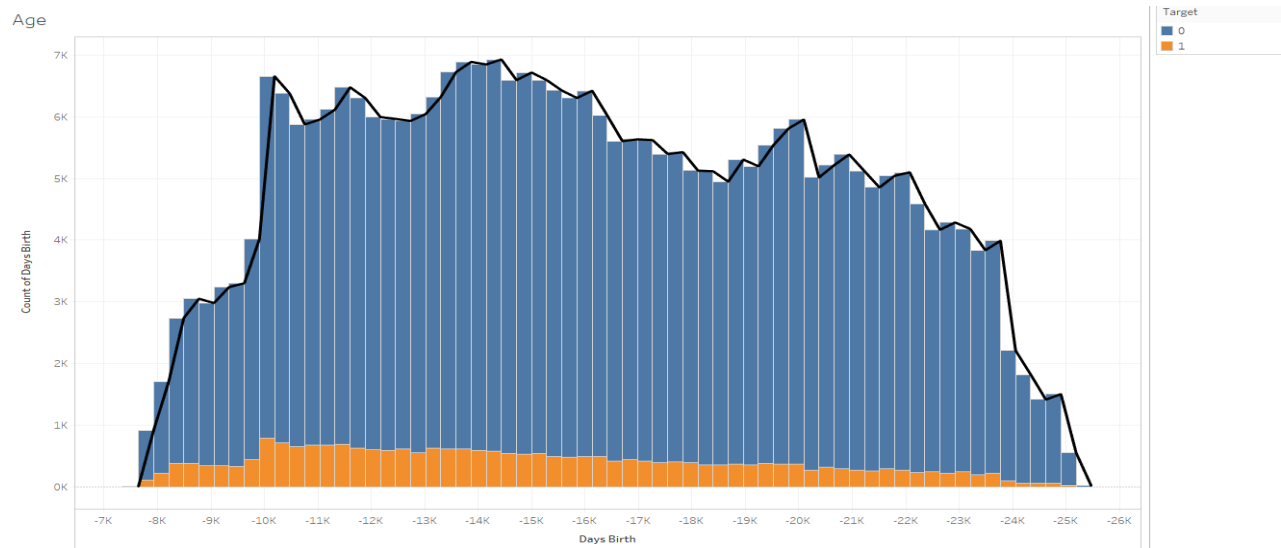


Most of the applicants who applied for loan had 2 family members in total

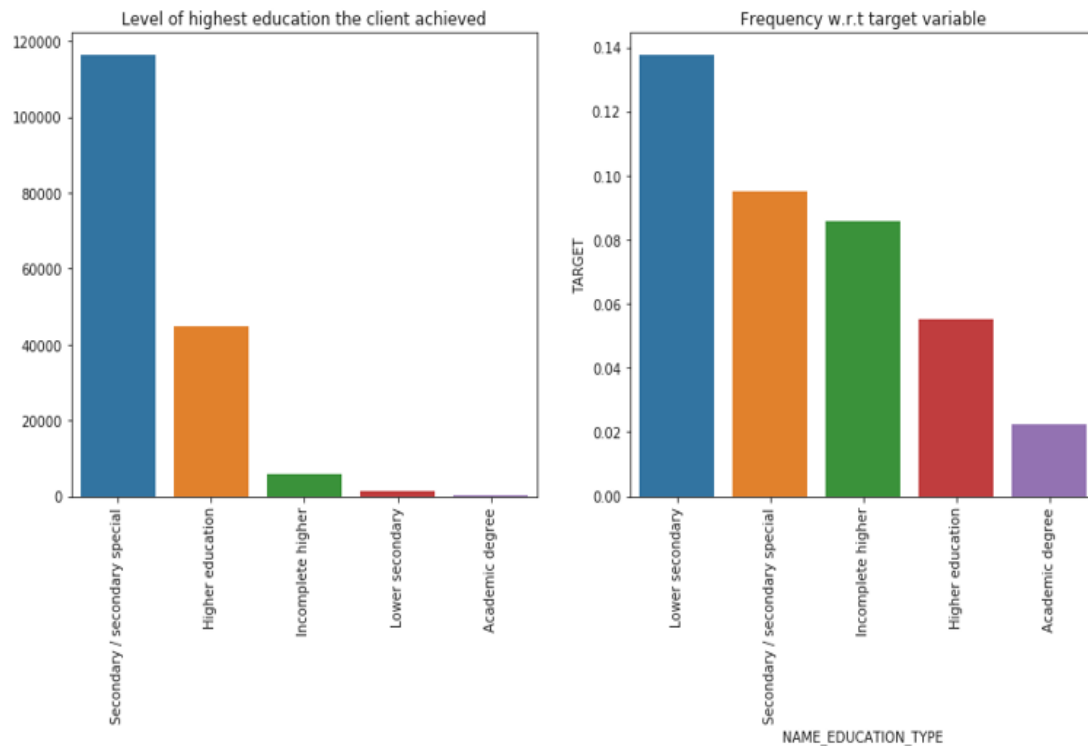
## DAYS\_BIRTH

The age range is between approximative 20 and 68 years.

When compared it with the target variable we can see that clients with payment difficulties are more concentrated in the age group 25-35 and it goes on decreasing as the age increases.

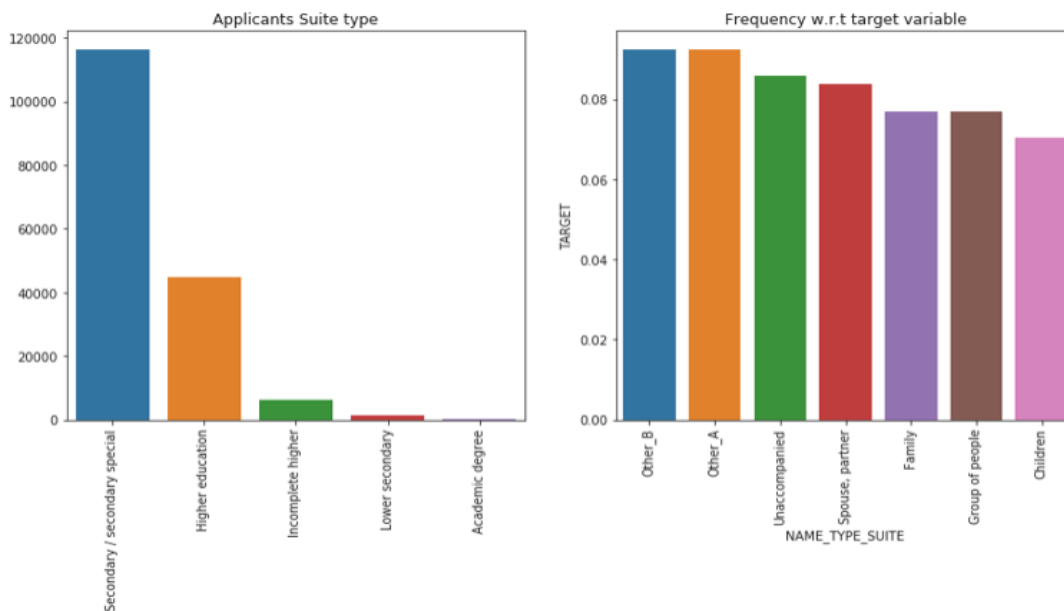


## NAME\_EDUCATION\_TYPE

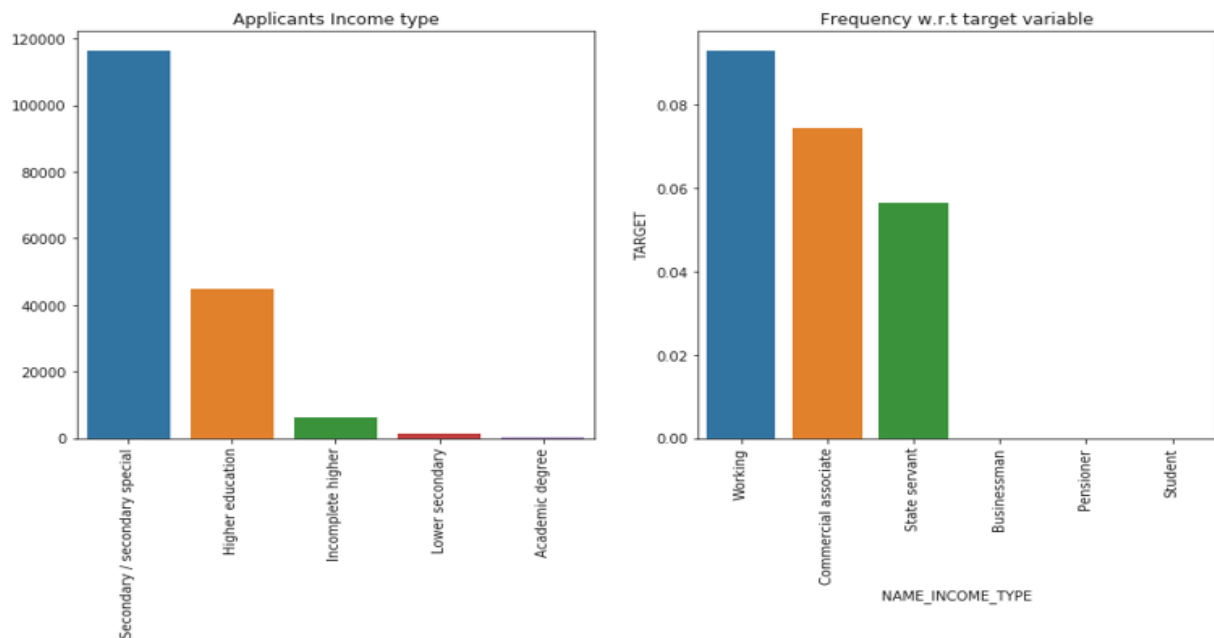


A large number of applications (120K) are filed by people having secondary education followed by people with Higher Education with 40K applications. we see that the applicants with Lower Secondary education status has the highest percentage of payment related problems.

## NAME\_TYPE\_SUITE

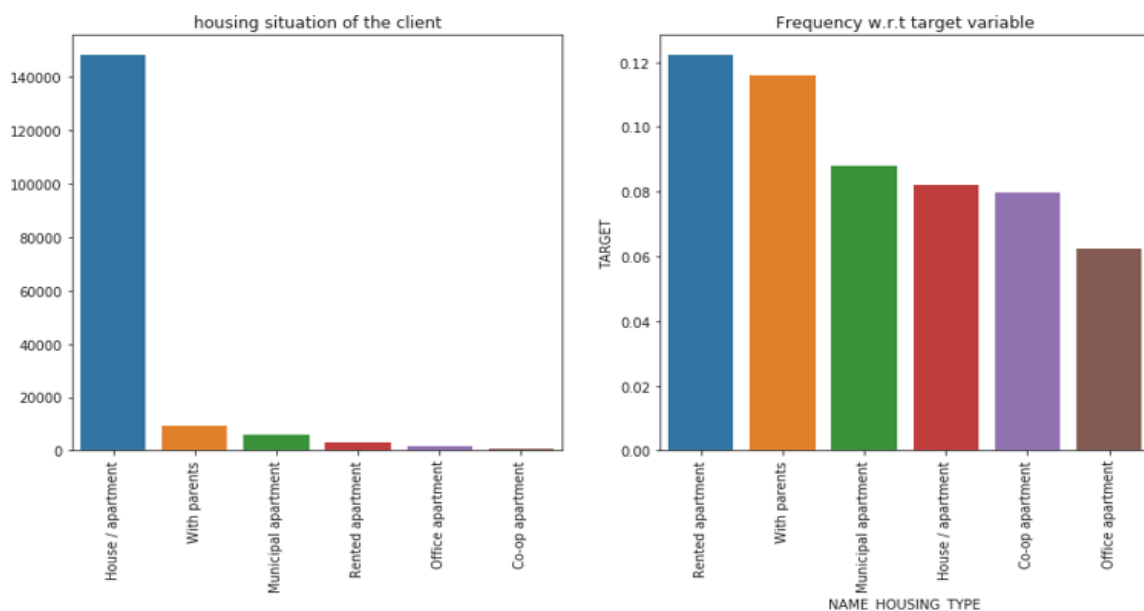


## NAME\_INCOME\_TYPE



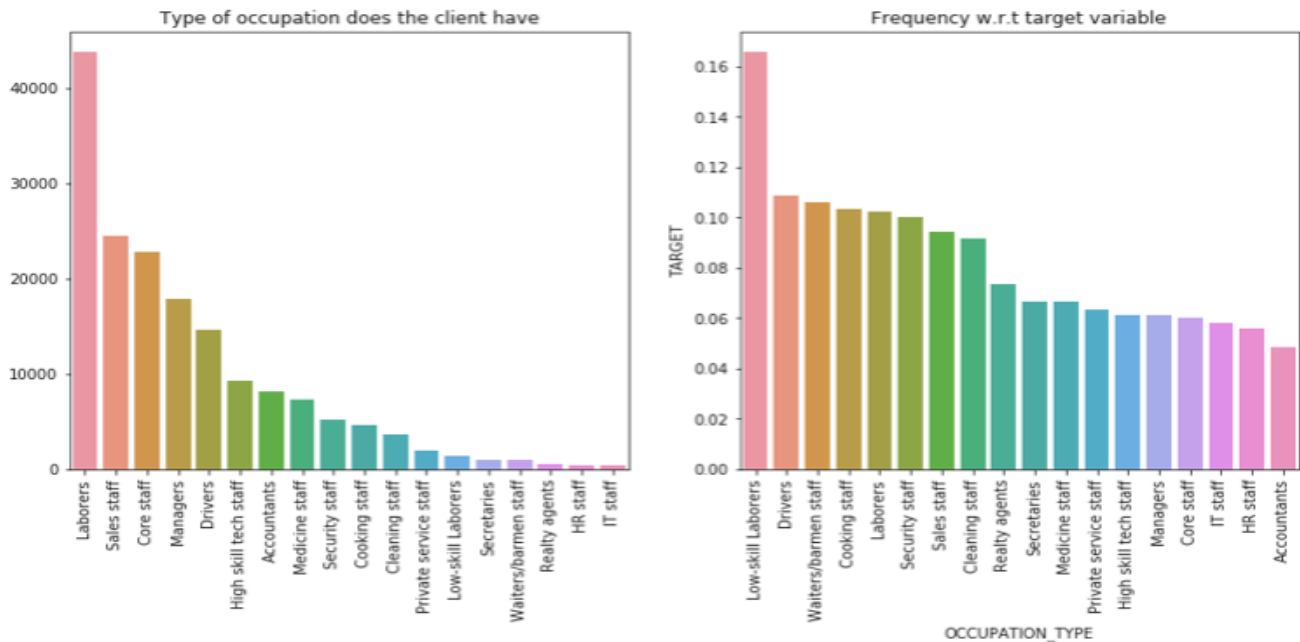
We see that Applicants having Income Types: Working & Commercial associate has the highest percentage (about 90% and 70% approx.) of Target = 1 i.e. having more payment problems, while Pensioners have the least (about 5.3%).

## NAME\_HOUSING\_TYPE



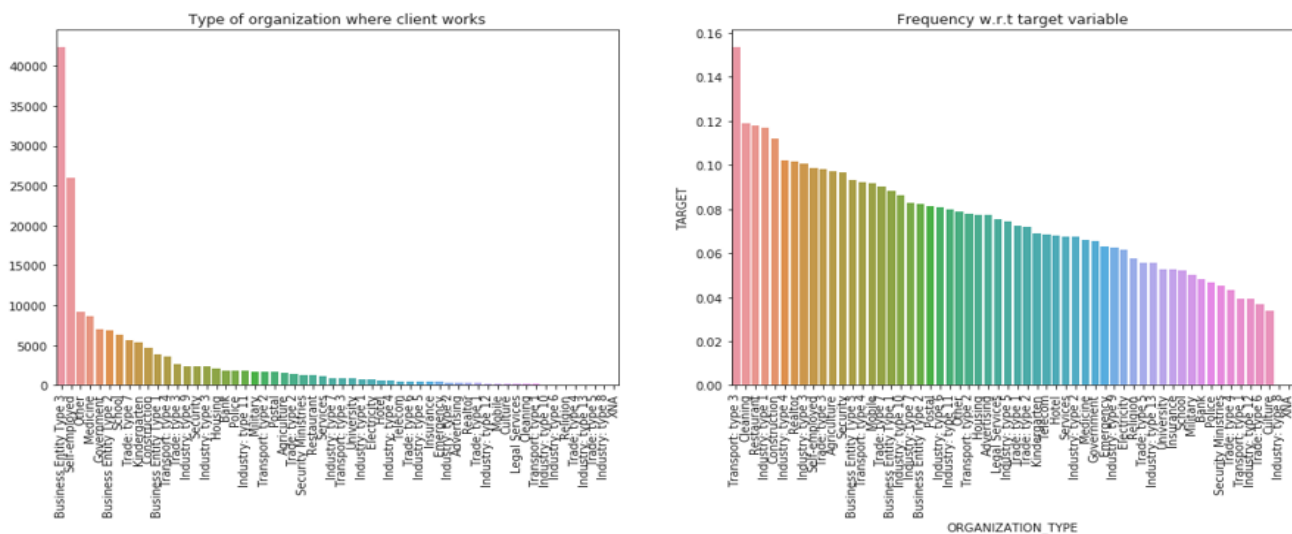
Applicants living in House / apartments has the highest number of loan applications equal to 140K. While we see that Rented apartment and applicants living with parents has highest percentage (about 12% and 11% approx.) of Target = 1 i.e. having more payment problems

## OCCUPATION\_TYPE



We can see that laborer's (45k), Sales ,core staff has highest number of loan applications but wen it comes to Non loan repayment highest number can be seen in Low-skill laborer's(16%),Drivers,Walters,cookin staff and so on...

## ORGANIZATION\_TYPE



Applicants works mostly in an organization of Business Entity Type 3(more than 40k), second number is of the Self-employed. But when it comes to Non loan repayment highest number can be seen in Applicants wo work in Transport Type3(~15%)

DAYS_EMPLOYED	0.071049	REGION_POPULATION_RELATIVE	-0.036097
DAYS_BIRTH	0.066953	AMT_CREDIT	-0.038297
REGION_RATING_CLIENT_W_CITY	0.061690	AMT_GOODS_PRICE	-0.047942
REGION_RATING_CLIENT	0.059035	EXT_SOURCE_2	-0.164536
DAYS_LAST_PHONE_CHANGE	0.056916	EXT_SOURCE_3	-0.180401

**Top most co-related (Positive & Negative) columns w.r.t target variables are as follows**

DAYS_EMPLOYED	REGION_POPULATION_RELATIVE
DAYS_BIRTH	AMT_CREDIT
REGION_RATING_CLIENT_W_CITY	AMT_GOODS_PRICE
REGION_RATING_CLIENT	EXT_SOURCE_2
DAYS_LAST_PHONE_CHANGE	EXT_SOURCE_3

## Previous Application Dataset

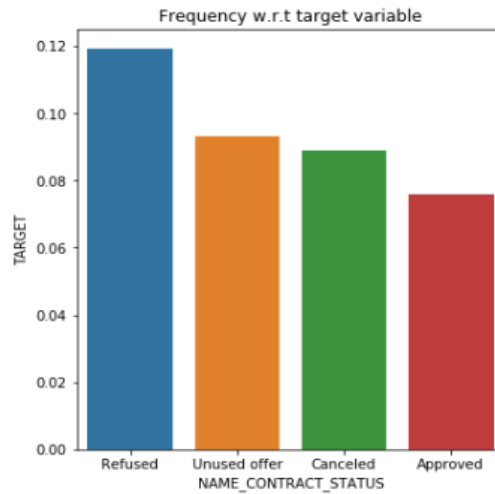
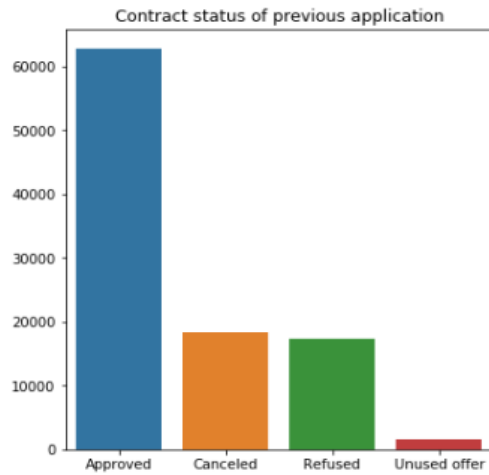
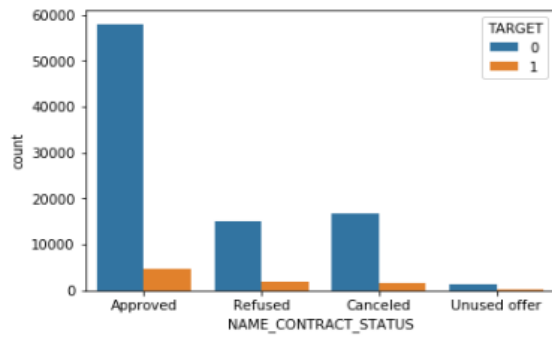
- Merge the datasets Application and previous on SK\_Current\_ID (inner join)
- Took sample out of the output data
- Started visualisation.

- **Performing bivariate analysis for NAME\_CONTRACT\_STATUS variable with the Target variable**

There are four types of contract\_status in the previous application data:  
Approved,Cancelled,refused,unused offer. Approved are almost triple the number around 60k.

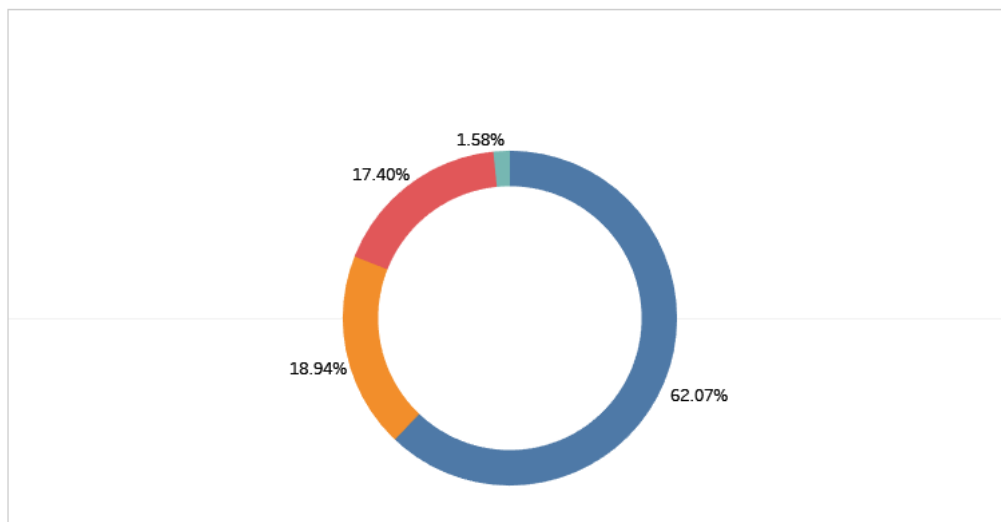
In terms of percent of defaults for current applications in the sample, clients with history of previous applications have largest percent of defaults when in their history contract statuses are Refused about 12%, followed by Unused offer, Canceled and Approved (lowest percent of defaults in current applications, with less than 8%).

Please find the below screenshot for reference



We have done the similar analysis in Tableau as well for the Contract\_status field

Contract status



Name Contract Status

- Approved
- Canceled
- Refused
- Unused offer

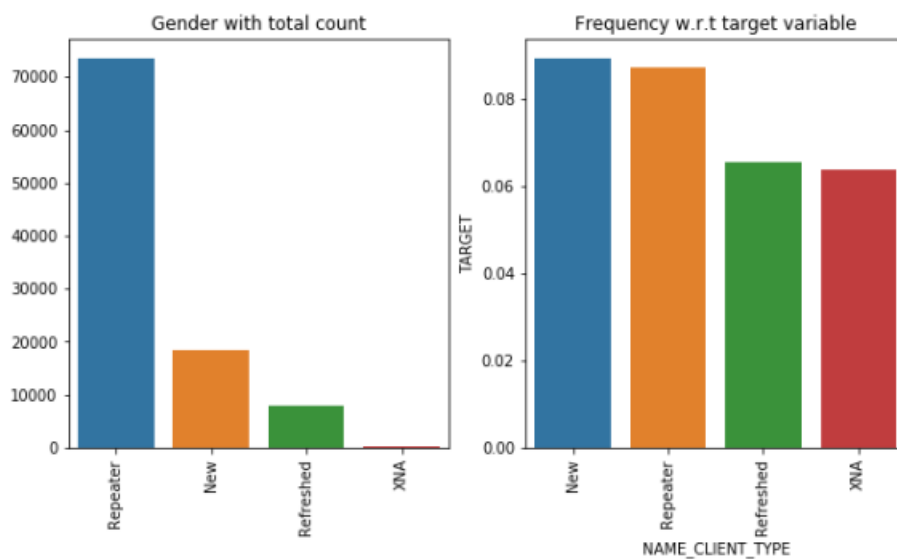
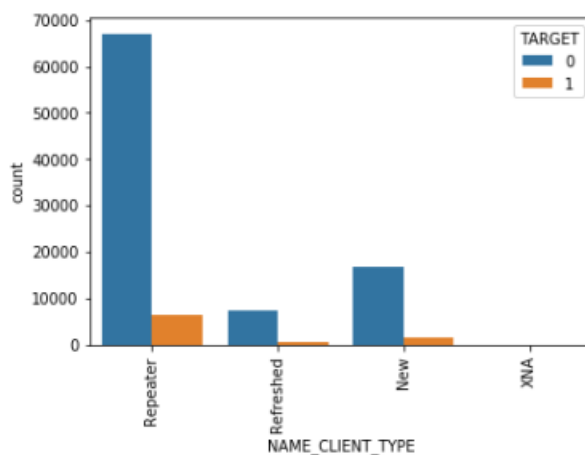
### Contract was approved or not in previous application:

- Approved: 62.1 % times
- Cancelled: 18.9 % times
- Refused: 17.4 % times
- Unused offer: 1.58 % times

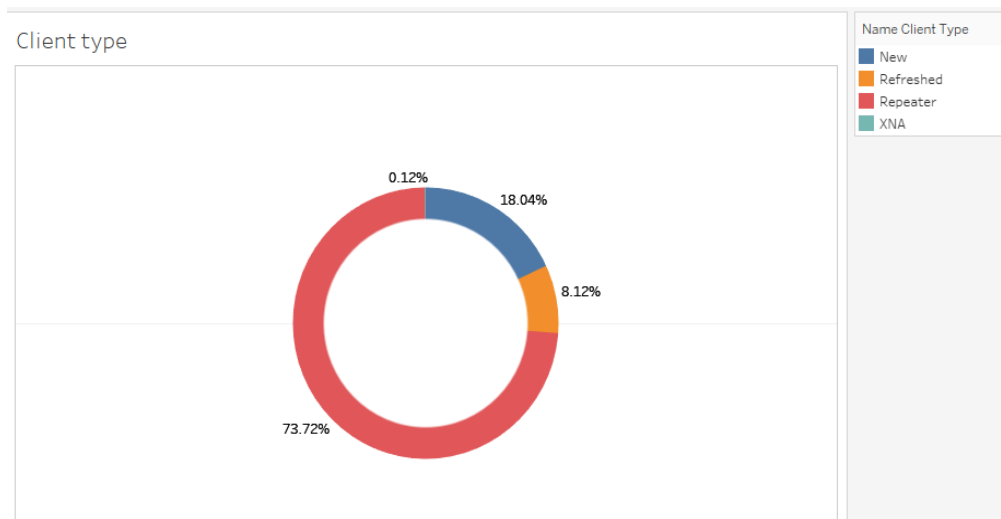
- **Performing bivariate analysis for NAME\_CLIENT\_TYPE variable with the Target variable**

As shown in the below plot , Most of the previous applications have client type Repeater (70k), just over 20K are New and ~10K are Refreshed.

In terms of default percent for current applications of clients with history of previous applications, current clients with previous applications have values of percent of defaults ranging from 8.5%, 8.25% and 7% corresponding to client types in the past New, Repeater and Refreshed, respectively.

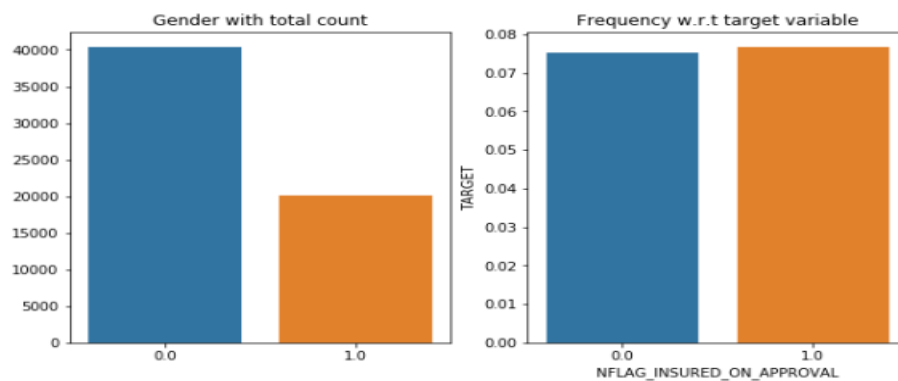
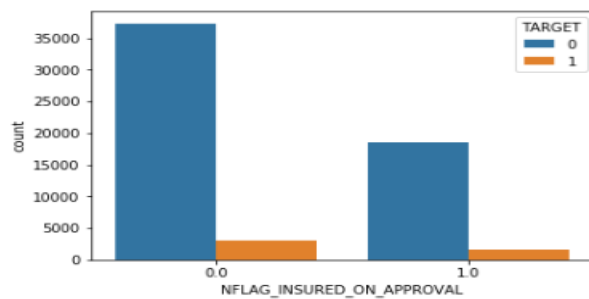


We have done the similar analysis in Tableau as well for the Client\_Typ field, Approximately 74 % was repeater clients who applied for previous application.



- Performing bivariate analysis for NFLAG\_INSURED\_ON\_APPROVAL variable with the Target variable

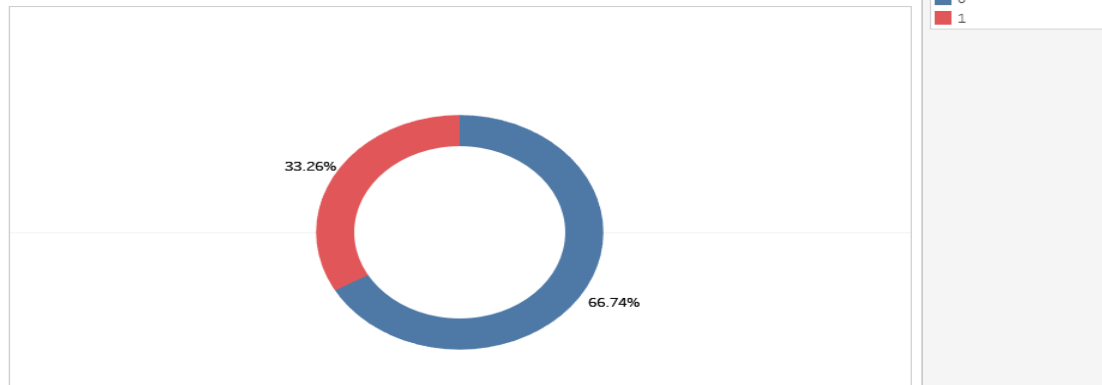
The frequency of the client with insurance is same, there no much difference weather client had insurance or not in the previous application



Did the client request insurance during the previous application? (Yes :1 or NO: 0)



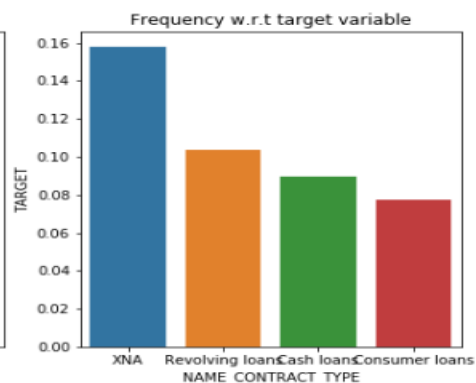
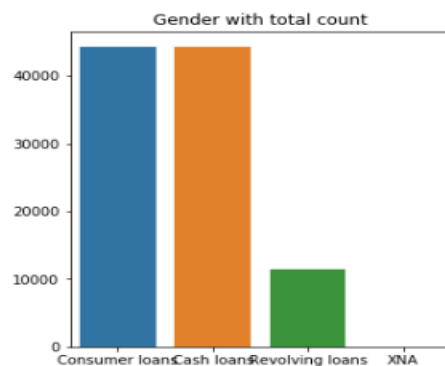
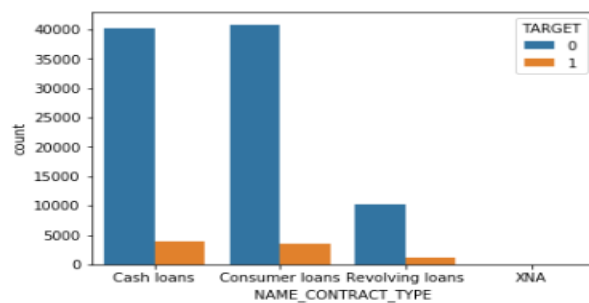
Insurance



- Performing bivariate analysis for **NAME\_CONTRACT\_TYPE** variable with the Target variable

As shown in the below screenshot, There are three types of contract in the previous application data: Cash loans, Consumer loans, Revolving loans. Cash loans and Consumer loans are almost the same number around 40k whilst Revolving loans are 15K.

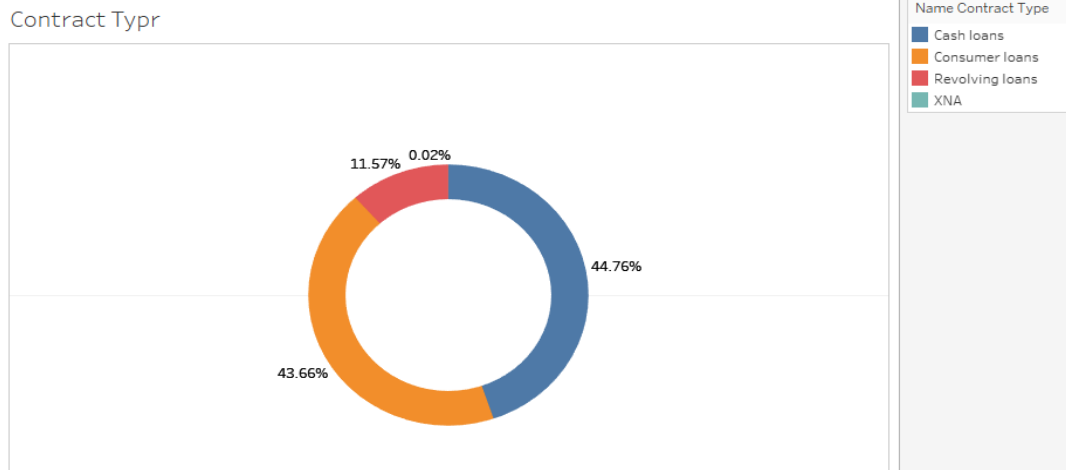
The percent of defaults loans for clients with previous applications is different for the type of previous applications contracts, decreasing from near about 10% for Revolving loans then about 9.5% for Cash loans and 8% for Consumer loans.



We have done the same analysis in Tableau for Contractor type variable,

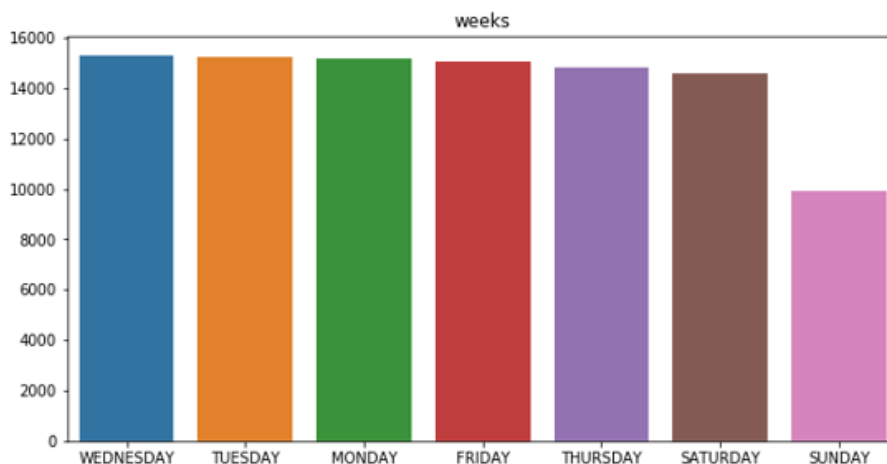
### Contract product type of previous application:

- Cash loans - 44.76 %
- Consumer loans - 43.66 %
- Revolving loan - 11.57 %
- XNA - 0.0207 %



- **Performing for univariate Analysis for WEEKDAY\_APPR\_PROCESS\_START variable**

It is strange that client applied only on weekdays for the Previous application. The number of previous application is same through the weekdays.



It is strange that client applied only on weekdays for the Previous application. The number of previous application is same through the weekdays.

Hence, we have made our Analysis on Credit using two datasets. We have successfully done analysis on Categorical as well as numerical variables, performed Univariate, Bivariate analysis and also found out Top 10 variables tat are Co-related with the Target variable

