# LOAN DEFAULT (CASE STUDY)

**DATA CLEANSING AND EXPLORATORY DATA ANALYSIS**

PROJECT BY:

MADHAVAN RANGARAJAN

# DATA USED

***2 files have been used for the project as explained below:***

- ***1. '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.***

- ***2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.***

Note:

FOR OUR CONVINIENCE HERE WE REFER CLIENT WITH PAYMENT DIFFICULTIES AS 'DEFAULTER' AND OTHER CLIENTS AS 'REPAYER' WHEN WE PERFORM SEGMENTATION BASED ON TARGET VARIABLE

# TOOLS AND LIBRARIES USED

1) The whole case-study was done using **python**
2) Libraries used for visualization
    A) matplotlib
    B) seaborn
3) Libraries used for analysis
    A) numpy
    B) pandas
    C) scipy.stats
    D) math

(PANDAS PROFILING LIBRARY UN-USED)

# MAIN APPLICATION DATA - DISCREPENCIES AND DATA CLEANING

**Lets inspect all the columns with greater than 40% missing values**

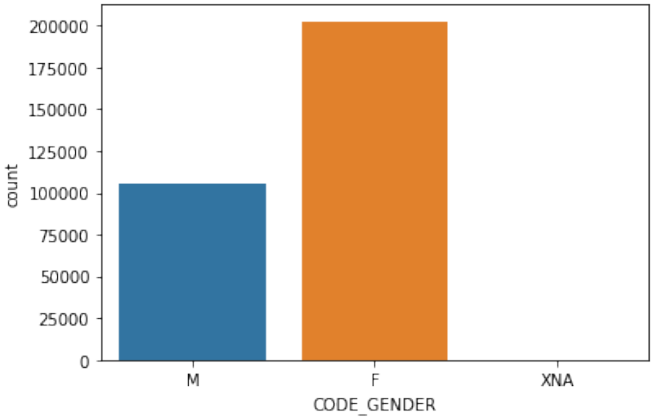
```
above_40 = missing[missing > 40]
above_40
```

```
COMMONAREA_MEDI                69.87
COMMONAREA_AVG                 69.87
COMMONAREA_MODE                69.87
NONLIVINGAPARTMENTS_MODE       69.43
NONLIVINGAPARTMENTS_MEDI       69.43
NONLIVINGAPARTMENTS_AVG        69.43
FONDKAPREMONT_MODE             68.39
LIVINGAPARTMENTS_MEDI          68.35
LIVINGAPARTMENTS_MODE          68.35
LIVINGAPARTMENTS_AVG           68.35
FLOORSMIN_AVG                  67.85
FLOORSMIN_MEDI                 67.85
FLOORSMIN_MODE                 67.85
YEARS_BUILD_MEDI               66.50
YEARS_BUILD_MODE               66.50
YEARS_BUILD_AVG                66.50
OWN_CAR_AGE                    65.99
LANDAREA_AVG                   59.38
LANDAREA_MEDI                  59.38
LANDAREA_MODE                  59.38
BASEMENTAREA_AVG               58.52
BASEMENTAREA_MODE              58.52
BASEMENTAREA_MEDI              58.52
EXT_SOURCE_1                   56.38
NONLIVINGAREA_MEDI             55.18
NONLIVINGAREA_AVG              55.18
NONLIVINGAREA_MODE             55.18
ELEVATORS_AVG                  53.30
ELEVATORS_MEDI                 53.30
```

**MOST OF THE COLUMNS WHICH HAS ABOVE 40% MISSING VALUES IS `HOUSING DATA` except 'OWN_CAR_AGE' and 'EXT_SOURCE_1'. Housing data has huge missing values. By imputing missing values we will bring bias. Also, we already have columns 'FLAG_OWN_REALTY', 'NAME_HOUSING_TYPE' which tells us the relation between 'target-variable' and housing. So , lets drop the HOUSING DATA**

```
apply_data['ORGANIZATION_TYPE'].value_counts()[:5]
```

```
Business Entity Type 3    67992
XNA                       55374
Self-employed             38412
Other                     16683
Medicine                  11193
Name: ORGANIZATION_TYPE, dtype: int64
```




**XNA and XAP actually are null objects which was mistreated. We replace such values with np.nan**

# MAIN APPLICATION DATA -
# DISCREPENCIES AND CLEANING

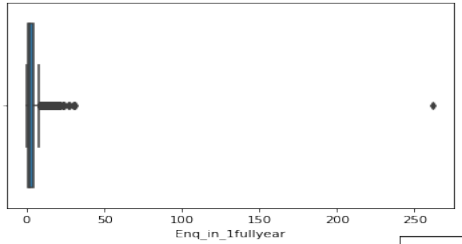
| | |
|---|---|
| AMT_REQ_CREDIT_BUREAU_HOUR | Number of enquiries to Credit Bureau about the client one hour before application |
| AMT_REQ_CREDIT_BUREAU_DAY | Number of enquiries to Credit Bureau about the client one day before application (excluding one hour before application) |
| AMT_REQ_CREDIT_BUREAU_WEEK | Number of enquiries to Credit Bureau about the client one week before application (excluding one day before application) |
| AMT_REQ_CREDIT_BUREAU_MON | Number of enquiries to Credit Bureau about the client one month before application (excluding one week before application) |
| AMT_REQ_CREDIT_BUREAU_QRT | Number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application) |
| AMT_REQ_CREDIT_BUREAU_YEAR | Number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application) |

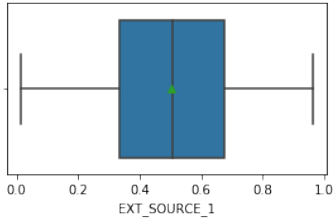
**Lets check the columns with more than 10% missing values as we have already treated for above 40% missing**

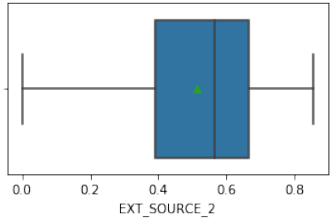
```
above_10 = missing[missing > 10]
above_10
```

```
EXT_SOURCE_1                56.38
OCCUPATION_TYPE             31.35
EXT_SOURCE_3                19.83
AMT_REQ_CREDIT_BUREAU_YEAR  13.50
AMT_REQ_CREDIT_BUREAU_MON   13.50
AMT_REQ_CREDIT_BUREAU_WEEK  13.50
AMT_REQ_CREDIT_BUREAU_DAY   13.50
AMT_REQ_CREDIT_BUREAU_HOUR  13.50
AMT_REQ_CREDIT_BUREAU_QRT   13.50
dtype: float64
```

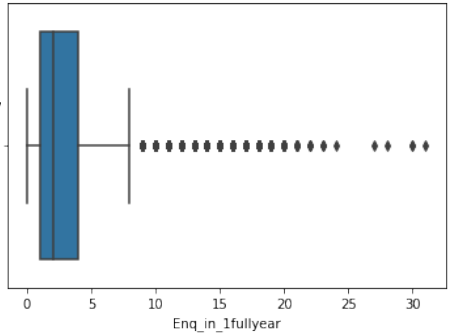
**So if we add all the rows we will get no of total enquires to credit bureau about the applicant in THE LAST ONE YEAR BEFORE APPLICATION . By summing all such columns we can analyise data without loss and at the same time reduce the complexity of the analysis**



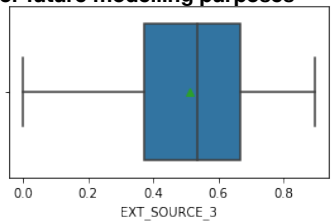



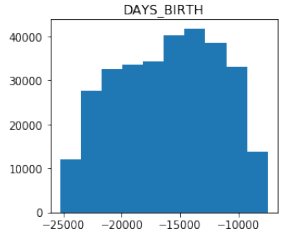

***Now after removing that row we can see that our boxplot is realistic***

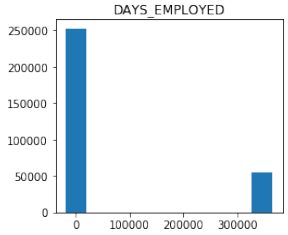


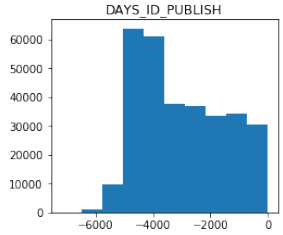

**Though data from Source 1 has 56.38 % missing values it is perfectly normally distributed . It may not be useful for EDA but it is very useful while developing models . So before lets b ackup it and save for future modelling purposes**

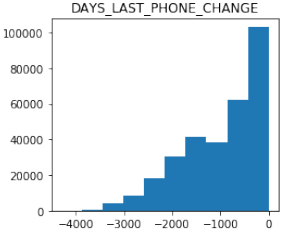


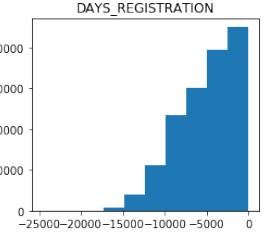
## MAIN APPLICATION DATA -
## DISCREPENCIES AND CLEANING



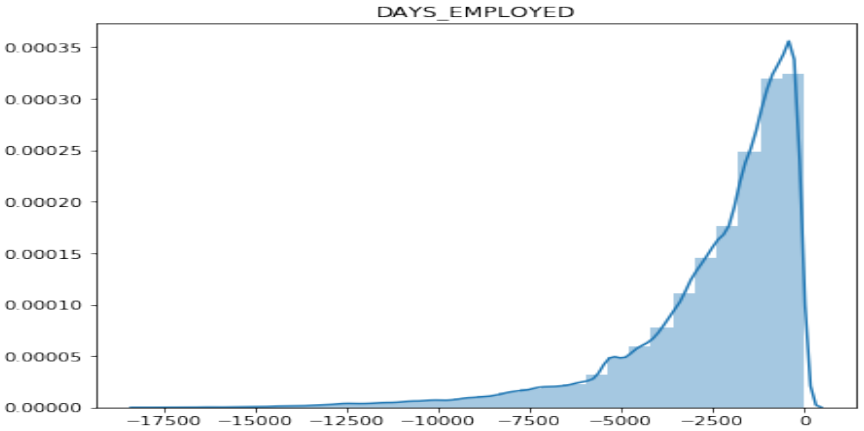




**All columns seem to be fine except 'Days_employed' which has positive values that too at a same point . 300000/365 = 821 years which is not possible . Lets inspect all values of 'DAYS_EMPLOYED' above 300000. For data above 300000 there is only one unique value which has repeated 55374 times. Lets replace that value to 'null'**

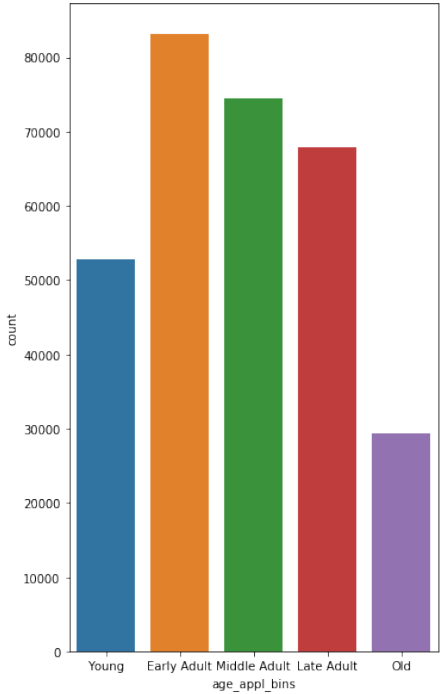


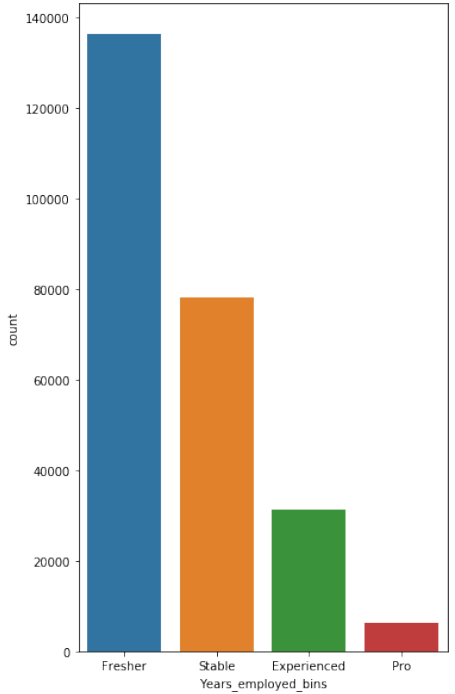

**Relative dates are huge numbers and analysis on such data is very tedious. So, lets convert it into years and positive figure s so that we can make interpretations on the data more easy.**

## BINNING DATA TO CONVERT IT INTO CATEGORICAL VARIABLE

- Early adults and Middle adults are people in ages 20-40.Freshers have less than 5 years of experience. We can hypothesize that greater proportion of people who claim loan start their careers late

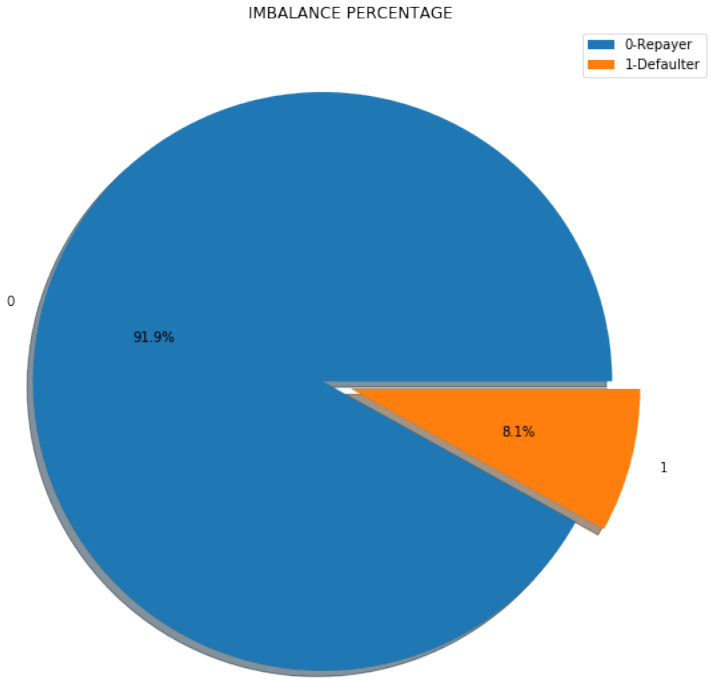
Count of people in each sub-category for AGE and EXPERIENCE respectively



# DATA IMBALANCE WITH RESPECT TO TARGET

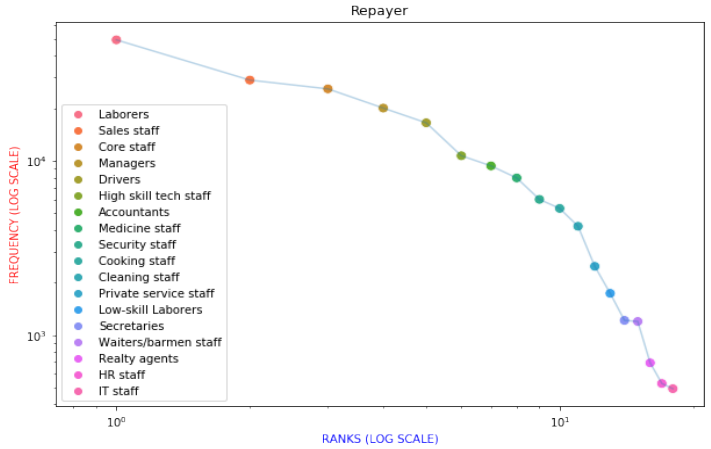
- THE DATA IS HIGHLY IMBALANCED
- FOR BUILDING MACHINE LEARNING MODELS WE NEED TO BALANCE THE DATA USING SAMPLING
- OUR MAIN GOAL HERE IS TO ANALYISE THE DATA AS A WHOLE AND COME UP WITH INTERPRETATIONS AND PATTERNS.
- THUS, WE CONTINUE OUR EDA

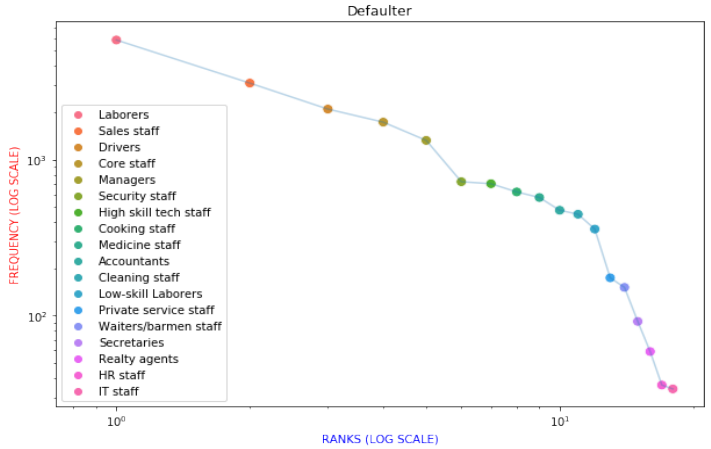


# UNI-VARIATE ANALYSIS ON CATEGORICAL DATA

Most of the object type columns have less sub - categories except Organizatio n type. When we use count or % bar plots to represent data with huge subcate gories it will be difficult to understand the distribution. Thus, as there is huge amount of data available lets plot a rank frequency distribution on a 'log-log' scale for both re payers and defaulters individually

RANK FREQUENCY DISTRIBUTION OF TYPE OF OCCUPATION



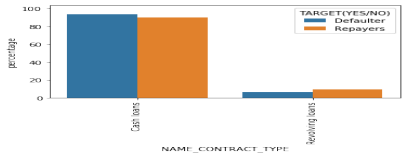



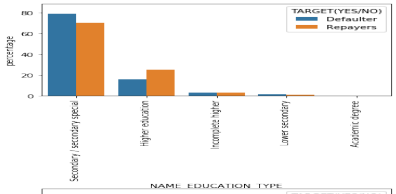

- Though both distributions look almost similar we can see a bump for 6-7 occupations in the lower middle end for Defaulter Data which indicates occupations ranging from Security staff to private service staff are occupations which are risk prone

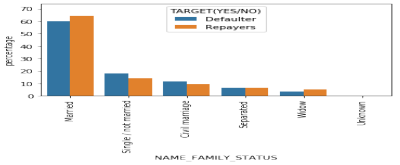
- Occupations like Security Staff, Cooking Staff , Low-skill laborers, Waiters, Accounts are occupations which involve risk in loan-approval

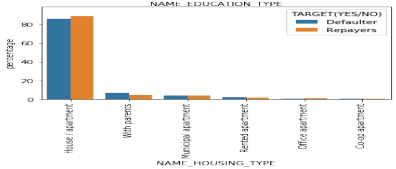
# UNI-VARIATE ANALYSIS ON CATEGORICAL DATA

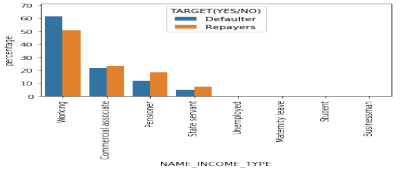
1) FEMALES ARE RELATIVELY RISK FREE THAN MEN
2) GREATER PROPORTION OF DEFAULTERS ARE MARRIED (MAY BE MIDDLE AGED)
3) SURPRSINGLY PRISONERS TEND TO REPAY LOAN MORE OFTEN
4) ALL LABOUR GROUPS ARE OCCUPATIONS WHERE HIGH RISK CAN BE SEEN
5) WE HAVE ALREADY HYPOTHESISED THAT PEOPLE WHO APPLY FOR LOAN TEND TO GET INTO EMPLOYMENT VERY LATE. SO YOUNG PEOPLE ARE MOSTLY UNEMPLOYED OR RECENT JOB HOLDERS. IT IS ALWAYS RISK APPROVING THEIR LOAN

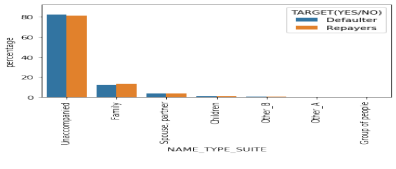



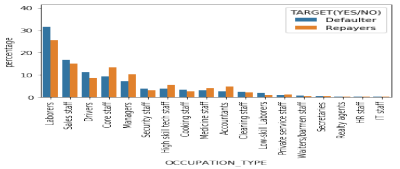



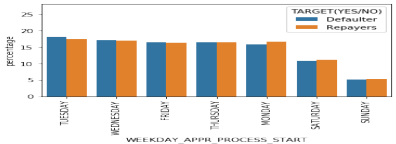



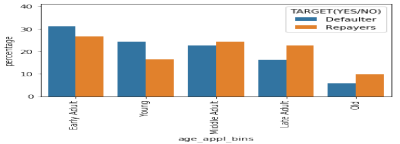



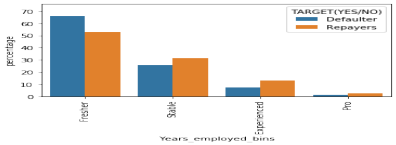



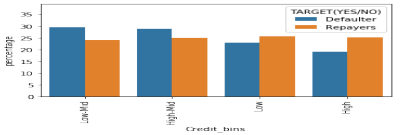



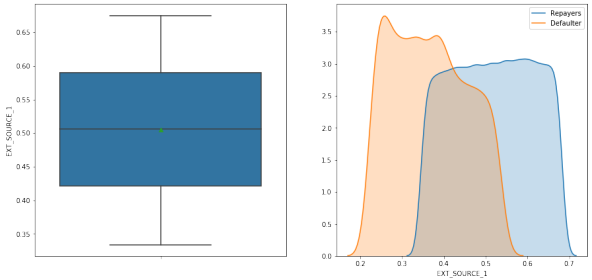



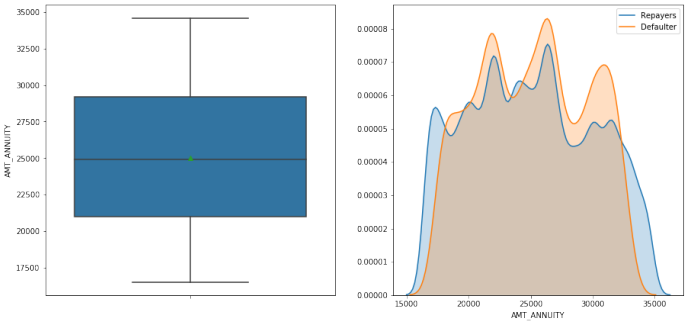



# SEGMENTED UNI-VARIATE ANALYSIS ON CATEGORICAL DATA

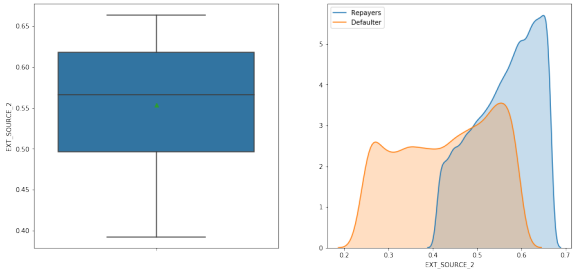
INTER QUARTILE RANGE(IQR) - Spread of EXT_SOURCE_1 for Total and Distribution of EXT_SOURCE_1 for (Repayers vs Defaulter)



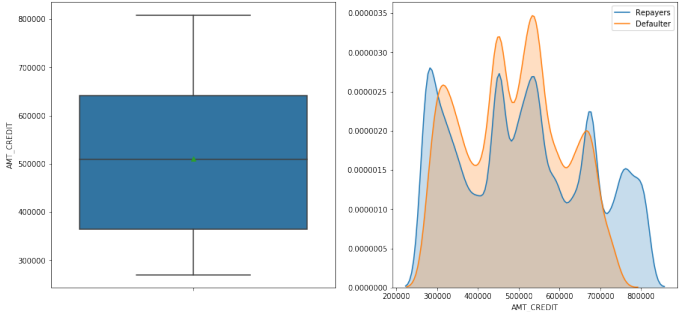

INTER QUARTILE RANGE(IQR) - Spread of AMT_ANNUITY for Total and Distribution of AMT_ANNUITY for (Repayers vs Defaulter)



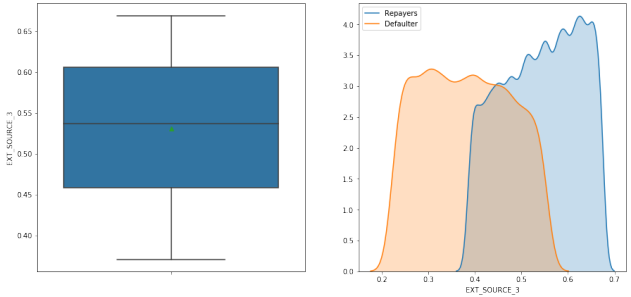

INTER QUARTILE RANGE(IQR) - Spread of EXT_SOURCE_2 for Total and Distribution of EXT_SOURCE_2 for (Repayers vs Defaulter)




INTER QUARTILE RANGE(IQR) - Spread of AMT_CREDIT for Total and Distribution of AMT_CREDIT for (Repayers vs Defaulter)




INTER QUARTILE RANGE(IQR) - Spread of EXT_SOURCE_3 for Total and Distribution of EXT_SOURCE_3 for (Repayers vs Defaulter)



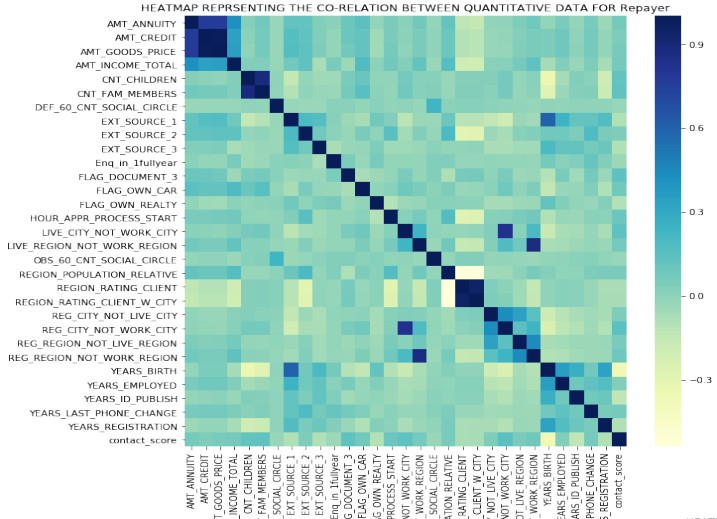

Inter quartile range is where major population lies. The distribution of Defaulters is more than Re payers in IQR. The bank should focus to shift the distribution of defaulters towards left side ( left skewed) . In other words for defaulters credit amount must be reduced to avoid risk

2) We can see that there are continuous spikes in the distribution of income which indicates there are some standard income packages followed across industries.

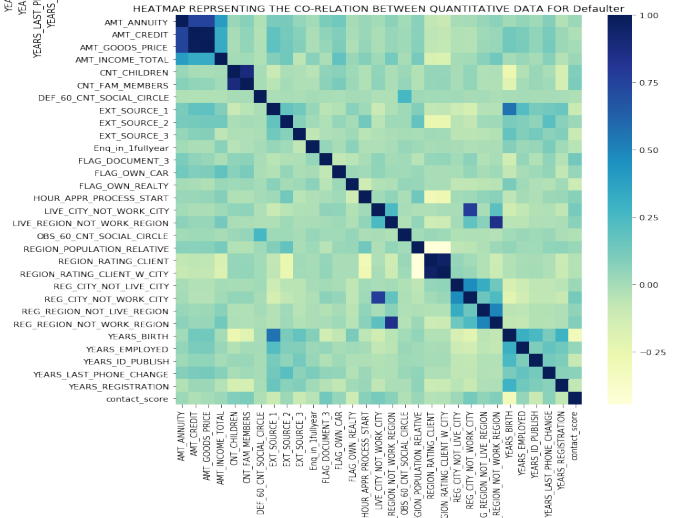
3) For all 3 credit score sources re payers inter quartile range is spread in higher credit scores wher eas defaulters inter quartile range is spread across lower credit scores. However among all 3 Ext source 1 seems more reliable because the distribution is perfectly normally distributed

# Bi-variate analysis after Segmentation on ‘Target’

| | Co-relation with TARGET |
|---|---|
| EXT_SOURCE_3 | -0.178899 |
| EXT_SOURCE_2 | -0.160454 |
| EXT_SOURCE_1 | -0.155321 |
| YEARS_BIRTH | -0.078229 |
| YEARS_EMPLOYED | -0.074952 |
| DEF_60_CNT_SOCIAL_CIRCLE | 0.031569 |
| LIVE_CITY_NOT_WORK_CITY | 0.032526 |
| FLAG_DOCUMENT_3 | 0.044340 |
| REG_CITY_NOT_LIVE_CITY | 0.044399 |
| REG_CITY_NOT_WORK_CITY | 0.051003 |




Most of the patterns are matching . There is not much variation across Def aulters and Re payers. But as there are many columns we need to observe patterns more closely (NUMERIC FORM)



# Bi-variate analysis after Segmentation on 'Target'

**Top 5 Positive Co-relations for Repayer Data**

| | | |
|---|---|---|
| AMT_GOODS_PRICE | AMT_CREDIT | 0.987250 |
| REGION_RATING_CLIENT_W_CITY | REGION_RATING_CLIENT | 0.950148 |
| CNT_FAM_MEMBERS | CNT_CHILDREN | 0.878571 |
| REG_REGION_NOT_WORK_REGION | LIVE_REGION_NOT_WORK_REGION | 0.861861 |
| REG_CITY_NOT_WORK_CITY | LIVE_CITY_NOT_WORK_CITY | 0.830379 |

**Top 5 Positive Co-relations for Defaulter Data**

| | | |
|---|---|---|
| AMT_CREDIT | AMT_GOODS_PRICE | 0.983103 |
| REGION_RATING_CLIENT | REGION_RATING_CLIENT_W_CITY | 0.956637 |
| CNT_FAM_MEMBERS | CNT_CHILDREN | 0.885481 |
| LIVE_REGION_NOT_WORK_REGION | REG_REGION_NOT_WORK_REGION | 0.847885 |
| REG_CITY_NOT_WORK_CITY | LIVE_CITY_NOT_WORK_CITY | 0.778537 |

ALL POSITIVE CO-RELATIONS ARE GENERAL AS WE KNOW THOSE VARIABLES ARE MEANT TO BE CO-RELATED. MUCH INTERESTING RELATION IS THE. RELATION BETWEEN CONTACT SCORE AND YEARS BIRTH . SO , AS PEOPLE WITH HIGHER AGE PROVIDE LESS CONTACT INFORMATION

]:

Region_rating_client has a negative co-relation with hour_appr_process_start
Which indicates the rating has considered geographic factors of the client
Location with respect to facilities available

**Top 5 Negative Co-relations for Repayer Data**

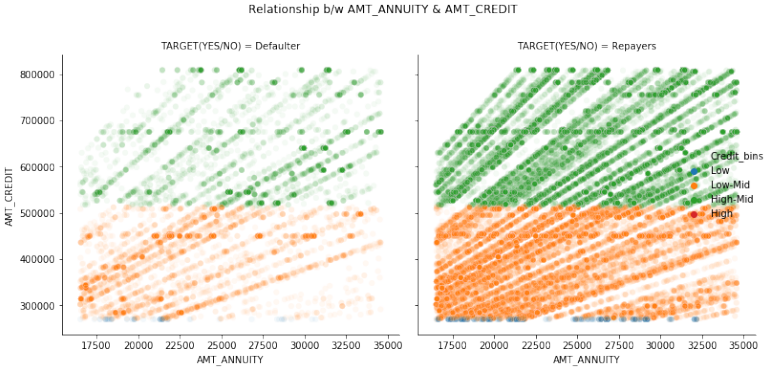
| | | |
|---|---|---|
| REGION_RATING_CLIENT | REGION_POPULATION_RELATIVE | -0.539008 |
| REGION_RATING_CLIENT_W_CITY | REGION_POPULATION_RELATIVE | -0.537305 |
| YEARS_BIRTH | contact_score | -0.359181 |
| CNT_CHILDREN | YEARS_BIRTH | -0.336910 |
| REGION_RATING_CLIENT | EXT_SOURCE_2 | -0.291619 |

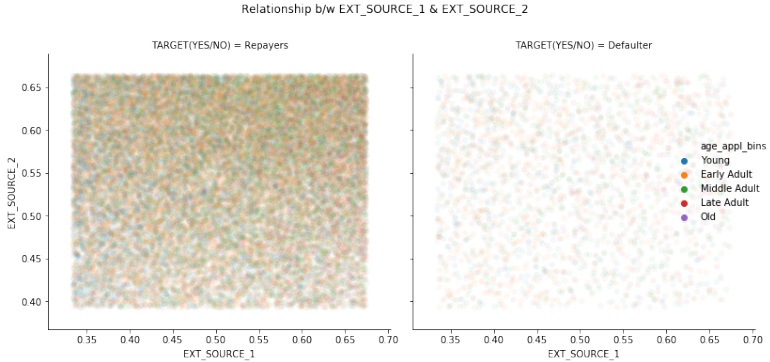
:

**Top 5 Negative Co-relations for Defaulter Data**

| | | |
|---|---|---|
| REGION_RATING_CLIENT_W_CITY | REGION_POPULATION_RELATIVE | -0.446990 |
| REGION_POPULATION_RELATIVE | REGION_RATING_CLIENT | -0.443250 |
| YEARS_BIRTH | contact_score | -0.307027 |
| HOUR_APPR_PROCESS_START | REGION_RATING_CLIENT | -0.293903 |
| REGION_RATING_CLIENT_W_CITY | HOUR_APPR_PROCESS_START | -0.275698 |

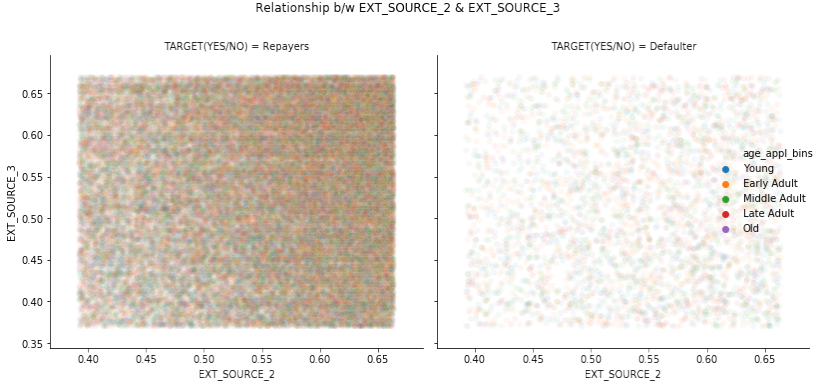
## Bi-variate analysis after Segmentation on ‘Target’




THE RELATIONSHIP BETWEEN EXT_SOURCE_2 AND EXT_SOURCE_3 seems very reliable source of judgement . The thickness on the top right block of the plot indicate that people with high scores from those 2 sources can be easily trusted upon. The bank should allot them loan

FOR BOTH REPAYER AND DEFAULTER CREDIT AMOUNT AND GOODS PRICE HAVE A GREAT POSITIVE RELATIONSHIP. HOWEVER WE CAN SEE SOMETHING INTERESTING. WHY BANKS APPROVING HIGHER LOAN AMOUNT MORE THAN THAT OF THE GOODS PRICE EVEN FOR. DEFAULTERS? TO AVOID RISK BANKS SHOULD ALWAYS MAINTAIN SLOPE OF THE SCATTER GREATER THAN 1

# PREVIOUS APPLICATION DATA: CLEANING AND MERGING

```
pre_ap.head()
```

| | SK_ID_PREV | SK_ID_CURR | NAME_CONTRACT_TYPE | AMT_ANNUITY | AMT_APPLICATION | AMT_CREDIT | AMT_DOWN_PAYMENT | AMT_GOODS_PRICE | WEEKD |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2030495 | 271877 | Consumer loans | 1730.430 | 17145.0 | 17145.0 | 0.0 | 17145.0 | |
| 1 | 2802425 | 108129 | Cash loans | 25188.615 | 607500.0 | 679671.0 | NaN | 607500.0 | |
| 2 | 2523466 | 122040 | Cash loans | 15060.735 | 112500.0 | 136444.5 | NaN | 112500.0 | |
| 3 | 2819243 | 176158 | Cash loans | 47041.335 | 450000.0 | 470790.0 | NaN | 450000.0 | |
| 4 | 1784265 | 202054 | Cash loans | 31924.395 | 337500.0 | 404055.0 | NaN | 337500.0 | |

## MISSING VALUES %

```
PRE_SK_ID_PREV                  0.00
PRE_SK_ID_CURR                  0.00
PRE_NAME_CONTRACT_TYPE          0.02
PRE_AMT_ANNUITY                22.29
PRE_AMT_APPLICATION             0.00
PRE_AMT_CREDIT                  0.00
PRE_AMT_DOWN_PAYMENT           53.64
PRE_AMT_GOODS_PRICE            23.08
PRE_WEEKDAY_APPR_PROCESS_START  0.00
PRE_HOUR_APPR_PROCESS_START     0.00
PRE_FLAG_LAST_APPL_PER_CONTRACT 0.00
PRE_NFLAG_LAST_APPL_IN_DAY      0.00
PRE_RATE_DOWN_PAYMENT          53.64
PRE_RATE_INTEREST_PRIMARY      99.64
PRE_RATE_INTEREST_PRIVILEGED   99.64
PRE_NAME_CASH_LOAN_PURPOSE     95.83
PRE_NAME_CONTRACT_STATUS        0.00
PRE_DAYS_DECISION               0.00
PRE_NAME_PAYMENT_TYPE          37.56
PRE_CODE_REJECT_REASON         81.33
PRE_NAME_TYPE_SUITE            49.12
PRE_NAME_CLIENT_TYPE            0.12
PRE_NAME_GOODS_CATEGORY        56.93
PRE_NAME_PORTFOLIO             22.29
PRE_NAME_PRODUCT_TYPE          63.68
PRE_CHANNEL_TYPE                0.00
PRE_SELLERPLACE_AREA            0.00
PRE_NAME_SELLER_INDUSTRY       51.23
PRE_CNT_PAYMENT                22.29
PRE_NAME_YIELD_GROUP           30.97
PRE_PRODUCT_COMBINATION         0.02
PRE_DAYS_FIRST_DRAWING         40.30
PRE_DAYS_FIRST_DUE             40.30
PRE_DAYS_LAST_DUE_1ST_VERSION  40.30
PRE_DAYS_LAST_DUE              40.30
PRE_DAYS_TERMINATION           40.30
PRE_NFLAG_INSURED_ON_APPROVAL  40.30
dtype: float64
```

**One very efficient way to extract all the information without loosing data while merging is using the measure s of central tendancy (i.e mean,median or mode). For ex: if a customer has applied for loan three times and claime d 10 rs , 20 rs , 30rs each time respectively we can group with respect to his current sk_id and extract its mean ,me dian or mode . But the question remains when to choose what?**

**1) When you want to group continuous data (ratio scale) we can use mean**

**2) As ordinal(ordered categorical) data cannot be represented with float median can be chosen**

**3) When you want to group categorical data mode can be chosen**

WE NEED:

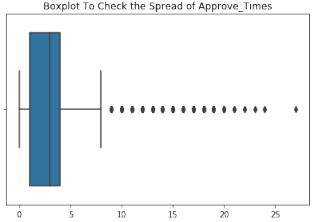
1) How many times loan has been approved?
2) How many times loan has been rejected?
3) What is the approval rate?
4) Was there any delay in the application process which made the client Cancel it?

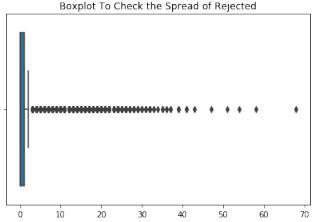
```
g = pre_ap.groupby('PRE_SK_ID_CURR')[['Approve_Times','Rejected','Applied_Times']].sum()
g.head()
```

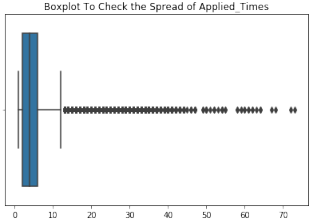
GROUPING WITH MEAN ON CONTINUOUS COLUMN AND MAKING NEW DERVIED COLUMNS TO GET APPROVAL %

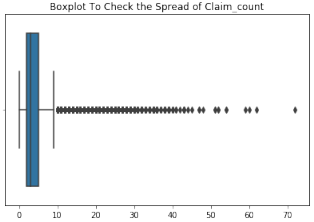
```
k = pre_ap.groupby('PRE_SK_ID_CURR')[allnum].mean()    #grouping continous columns with respect to 'SK_ID'
```

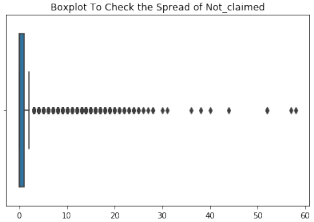
# PREV APPLICATION DATA. : UNIVARIATE AND SEGMENTED UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLES



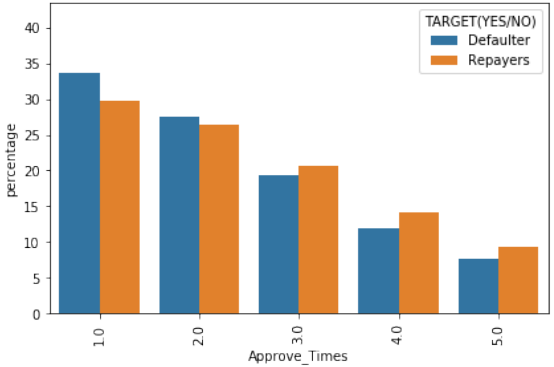



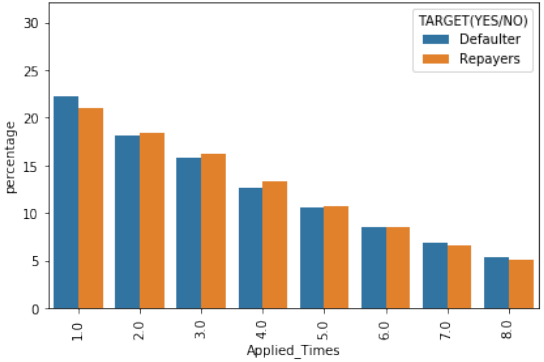




All these distributions are immensely left skewed

% CONTRIBUTION OF APPROVE_TIMES FOR REPAYER(LEFT) AND DEFAULTER(RIGHT) - (AFTER REMOVAL OF OUTLIERS)



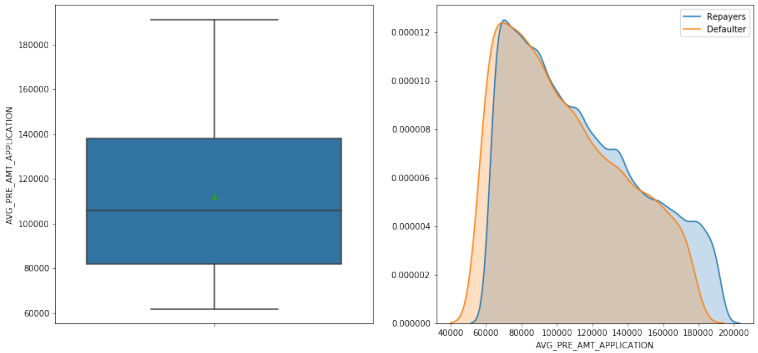

% CONTRIBUTION OF APPLIED_TIMES FOR REPAYER(LEFT) AND DEFAULTER(RIGHT) - (AFTER REMOVAL OF OUTLIERS)



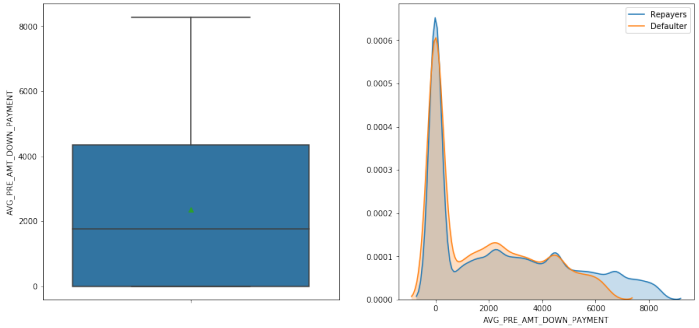

The distribution looks promising for both Defaulters and Re payers in these categories . But the bank should reduce the percentage of approval for defaulters for the first and second time. Atleast their loan approval duration must take longer time . Lets see it

# PREVIOUS APPLICATION DATA:
# UNIVARIATE AND SEGMENTED UNIVARIATE ANALYSIS ON CONTINUOUS VARIABLES

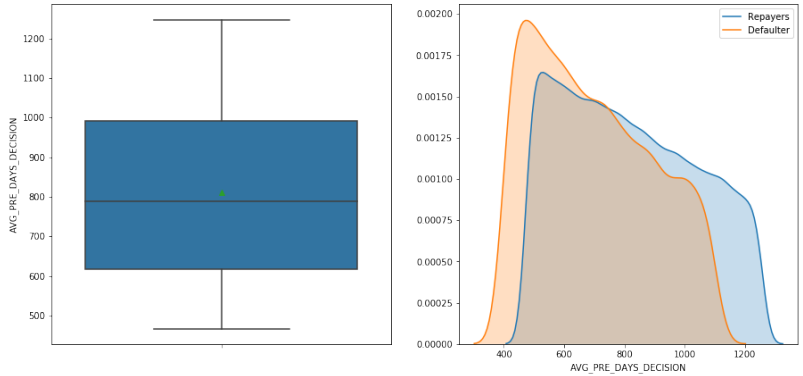
INTER QUARTILE(IQR) - Spread of AMT APPLICATION of Previous Application for Total and Distribution of AVG_PRE_AMT_APPLICATION for (Repayers vs Defaulter)




INTER QUARTILE(IQR) - Spread of AMT DOWN PAYMENT of Previous Application for Total and Distribution of AVG_PRE_AMT_DOWN_PAYMENT for (Repayers vs Defaulter)



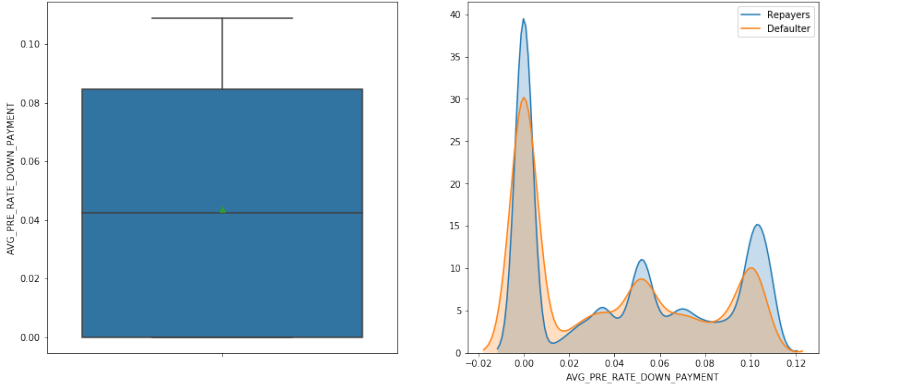

INTER QUARTILE(IQR) - Spread of DAYS DECISION of Previous Application for Total and Distribution of AVG_PRE_DAYS_DECISION for (Repayers vs Defaulter)




1) AVG_PRE_DAYS_DECISION typically indicates how fast or slow a person takes to apply for the next loan. In this context we can see that Defaulters tend to apply loan sooner after a previous loan has already been taken. The b ank should note risk in such type of customers who approve loan at a faster rate

2) REPAYERS HAVE BEEN CHARGED WITH HIGHER RATE OF DOWN PAYMENT . THE BANK SHOULD FOCUS ON GIVING LOAN ON A LESSER RATE OF DOWN PAYMENT AS LESS RISK IS INVOLVED

INTER QUARTILE(IQR) - Spread of RATE DOWN PAYMENT of Previous Application for Total and Distribution of AVG_PRE_RATE_DOWN_PAYMENT for (Repayers vs Defaulter)



# Co – relations in the variables belonging To previous application data

**Top 10 Co-relations for Repayer Data(WITH PREVIOUS APP DATA)**

| CO-RELATION | |
|---|---|
| Applied_Times and Enq_in_1fullyear | 0.553443 |
| Not_claimed and Enq_in_1fullyear | 0.513840 |
| Claim_count and Enq_in_1fullyear | 0.447961 |
| Approve_Times and Enq_in_1fullyear | 0.417606 |
| AVG_PRE_DAYS_DECISION and YEARS_LAST_PHONE_CHANGE | 0.337568 |
| AVG_PRE_AMT_ANNUITY and AMT_INCOME_TOTAL | 0.308243 |
| AVG_PRE_AMT_ANNUITY and AMT_ANNUITY | 0.290091 |
| Rejected and Enq_in_1fullyear | 0.289994 |
| AVG_PRE_AMT_GOODS_PRICE and Enq_in_1fullyear | 0.282397 |
| AVG_PRE_AMT_GOODS_PRICE and AMT_INCOME_TOTAL | 0.257872 |
| AVG_PRE_AMT_APPLICATION and AMT_INCOME_TOTAL | 0.245126 |

```
dict_corr['r for Defaulter']
```

**Top 10 Co-relations for Defaulter Data(WITH PREVIOUS APP DATA)**

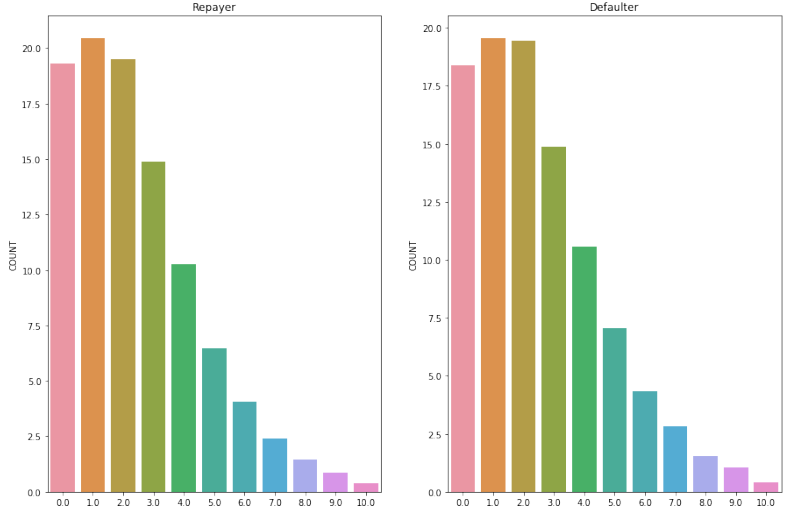
| CO-RELATION | |
|---|---|
| Applied_Times and Enq_in_1fullyear | 0.540146 |
| Not_claimed and Enq_in_1fullyear | 0.502669 |
| Claim_count and Enq_in_1fullyear | 0.439452 |
| Approve_Times and Enq_in_1fullyear | 0.402230 |
| AVG_PRE_AMT_ANNUITY and AMT_INCOME_TOTAL | 0.314716 |
| Rejected and Enq_in_1fullyear | 0.309692 |
| AVG_PRE_DAYS_DECISION and YEARS_LAST_PHONE_CHANGE | 0.303036 |
| AVG_PRE_AMT_GOODS_PRICE and Enq_in_1fullyear | 0.297294 |
| AVG_PRE_AMT_GOODS_PRICE and AMT_INCOME_TOTAL | 0.258038 |
| AVG_PRE_AMT_ANNUITY and Enq_in_1fullyear | 0.247496 |
| AVG_PRE_AMT_APPLICATION and AMT_INCOME_TOTAL | 0.242259 |

CO- RELATION

| | TARGET |
|---|---|
| Approval_Rate | 0.080392 |
| Rejected | 0.064477 |
| AVG_PRE_DAYS_DECISION | 0.046868 |
| AVG_PRE_AMT_ANNUITY | 0.034860 |
| AVG_PRE_RATE_DOWN_PAYMENT | 0.033597 |
| Approve_Times | 0.031549 |
| AVG_PRE_AMT_DOWN_PAYMENT | 0.024621 |
| AVG_PRE_AMT_APPLICATION | 0.021795 |
| Applied_Times | 0.019767 |
| Not_claimed | 0.018748 |
| AVG_PRE_AMT_CREDIT | 0.016106 |
| AVG_PRE_AMT_GOODS_PRICE | 0.015838 |
| Claim_count | 0.015804 |

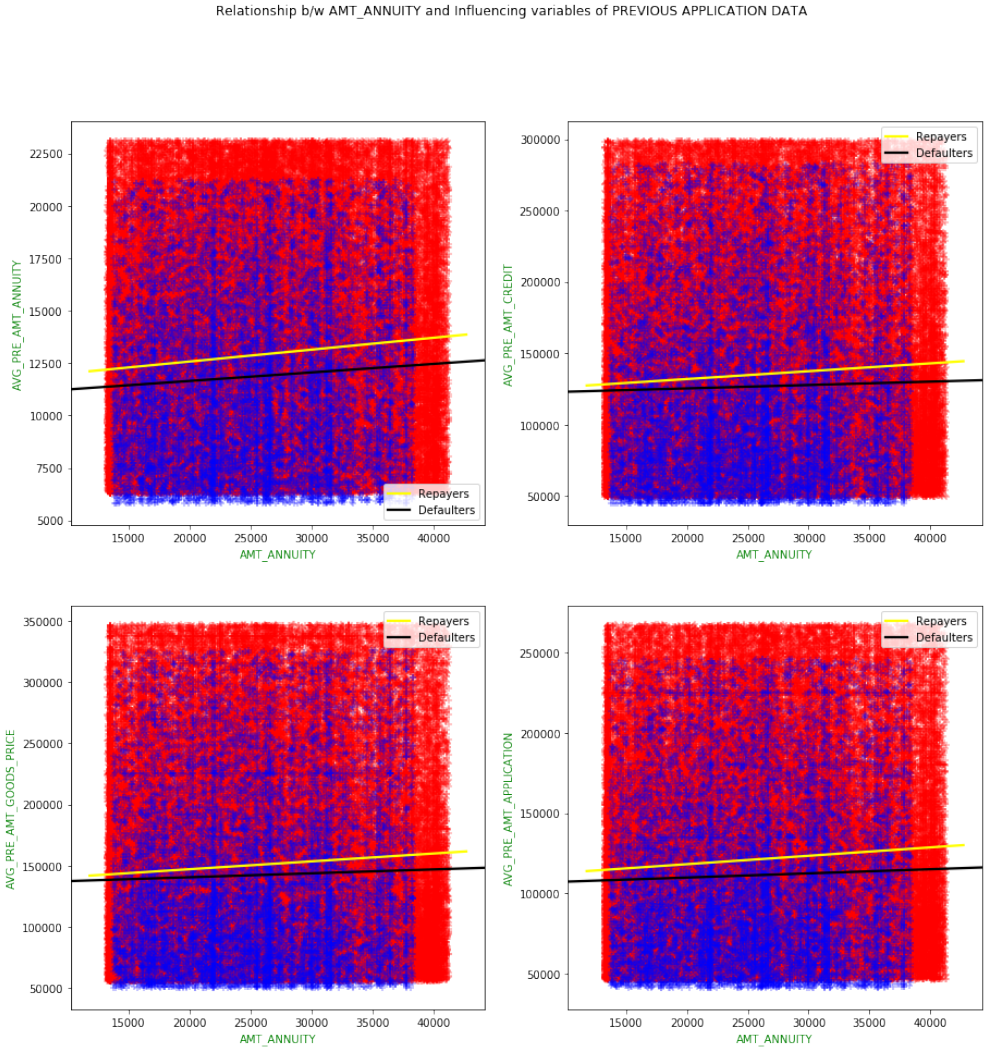
MANY OF THEM ARE CO RELATED WITH ENQUIRIES TO CREDIT BUREAU IN 1 FULL YEAR.(ENQ_IN_1fullyear ). This is very much interesting pattern. But all those columns of 'APPROVAL' data is co-related with ENQUIRIES because those columns are already highly co-related with each other . They form a cluster together.

% distribution ENQUIRIES TO CREDIT BUREAU IN 1 FULL YEAR

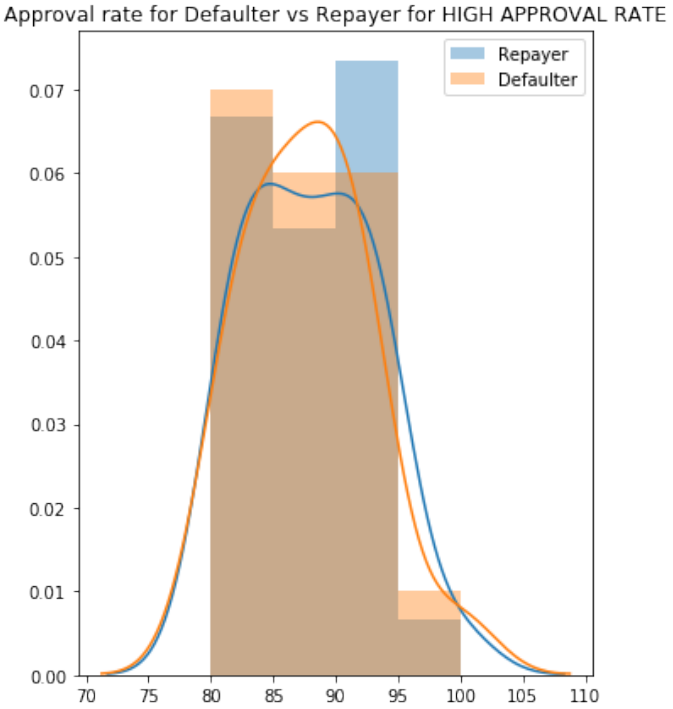


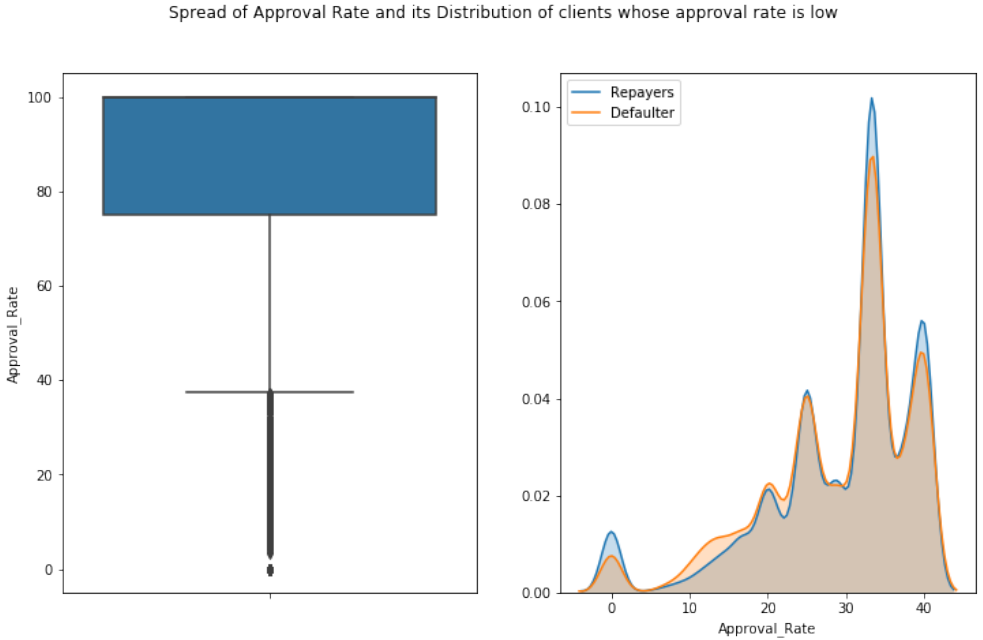
# Bi-variate analysis on influencing variables of Previous application data

- 1) The trend line for both Re payers and Defaulters follow a similar pattern in almost all the plots
- It is important to note the fact that for Re payers the slope is steeper than that of Defaulters
- This pattern will be very helpful in prediction of loan default when plotted for another small sample



# OTHER INTERESTING OBSERVATIONS




Repayers are having a lower approval rate and at the same time surprisingly the distribution of Defaulters are higher than Repayers for high approval rate. BANK SHOULD VERY MUCH FOCUS ON REDUCING THE APPROVAL RATES FOR DEFAULTERS AS HIGH RISK IS INVOLVED