CREDIT EDA CASE STUDY

SUBMITTED BY:

CHHAVI KANSAL

MUKESH CHAURASIA

SUMMARY

- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments.
- This will ensure that the consumers capable of repaying the loan are not rejected.
- Here, we need to explore the driving factors which are strong indicators of loan defaults.
- Two data sets were provided as part of this case study
  - Application Data
  - Previous application data

APPROACH

- Read the data set and check data type of each columns.
- Check all the columns with missing values and write approach to impute missing values.
- Drop the columns which has more than 50% missing values.
- Check datatype of all the columns and change the data type if needed.
- Check columns with outliers and explaining why they are outliers.
- Binning the continuous variable into categorical variable.
- Check columns with data imbalance.
- Differentiate between categorical columns and numerical columns and perform univariate and multivariate analysis against target variable on both application data and previous application data.
- Find out top 10 correlation in application data against target variable.
- At the end find top 3 variables and their correlation.

MISSING VALUES CHECK AND HANDLING

- Find the missing percentage of each column.
- Make two list of that data -
  - Having missing data greater than 50% (missing_column_50)
  - Having missing data lesser than 13 % (missing_column_13)
- Drop all columns which is having more than 50% missing data
- Do analysis for replacing the missing value for all those columns for less than 13% missing data. (Note: here we are not imputing the missing data value only representing the process).

87	HOUSETYPE_MODE	50.176091
65	FLOORSMAX_MODE	49.760822
79	FLOORSMAX_MEDI	49.760822
51	FLOORSMAX_AVG	49.760822
60	YEARS_BEGINEXPLUATATION_MODE	48.781019
74	YEARS_BEGINEXPLUATATION_MEDI	48.781019
46	YEARS_BEGINEXPLUATATION_AVG	48.781019
88	TOTALAREA_MODE	48.268517
90	EMERGENCYSTATE_MODE	47.398304
28	OCCUPATION_TYPE	31.345545
43	EXT_SOURCE_3	19.825307
116	AMT_REQ_CREDIT_BUREAU_HOUR	13.501631

DATA TYPE CORRECTION:

- Find the info of each column.
- Count the unique value of each column.
- Any column which has less than 10 unique values, we can consider these columns as categorical columns and change their data type as object.
- There are a total 40 column values as numeric, we converted all these columns into an object.

OUTLIER

- Here in our analysis to find out the outliers, we have considered numerical columns and analyzed the statistics of them.
- If we observe the application data columns, there are many columns with outlier values which are having a huge difference compared to the regular intervals of other values.
- We also designed the box plot or scatter plot to identify the outliers in below columns.
  - CNT_CHILDREN
  - AMT_INCOME_TOTAL
  - DAYS_EMPLOYED
  - OBS_30_CNT_SOCIAL_CIRCLE
  - AMT_REQ_CREDIT_BUREAU_MON

BINNING

- We checked continuous variable in data set and bin for better visualization and analysis
- After exploring all the continuous variables we have found below the column can be binned together.
  - AMT_INCOME_TOTAL
  - DAYS_BIRTH
  - CNT_FAM_MEMBERS
  - FLOORSMAX_AVG
  - TOTALAREA_MODE
  - FLOORSMAX_MEDI

AMT_REQ_CREDIT_BUREAU_YEAR	AMT_INCOME_TOTAL_cat	Age
1.0	avg	18-30
0.0	avg	40-50
0.0	low	50-60
NaN	avg	50-60
0.0	avg	50-60
...	...	...
NaN	avg	18-30
NaN	low	50-60
1.0	avg	40-50
0.0	avg	30-40
1.0	avg	40-50

IMBALANCE DATA

- We checked the TARGET variable for data imbalance and found there is data imbalance in the data.
- With below image it is clearly seen that we get imbalance data records.

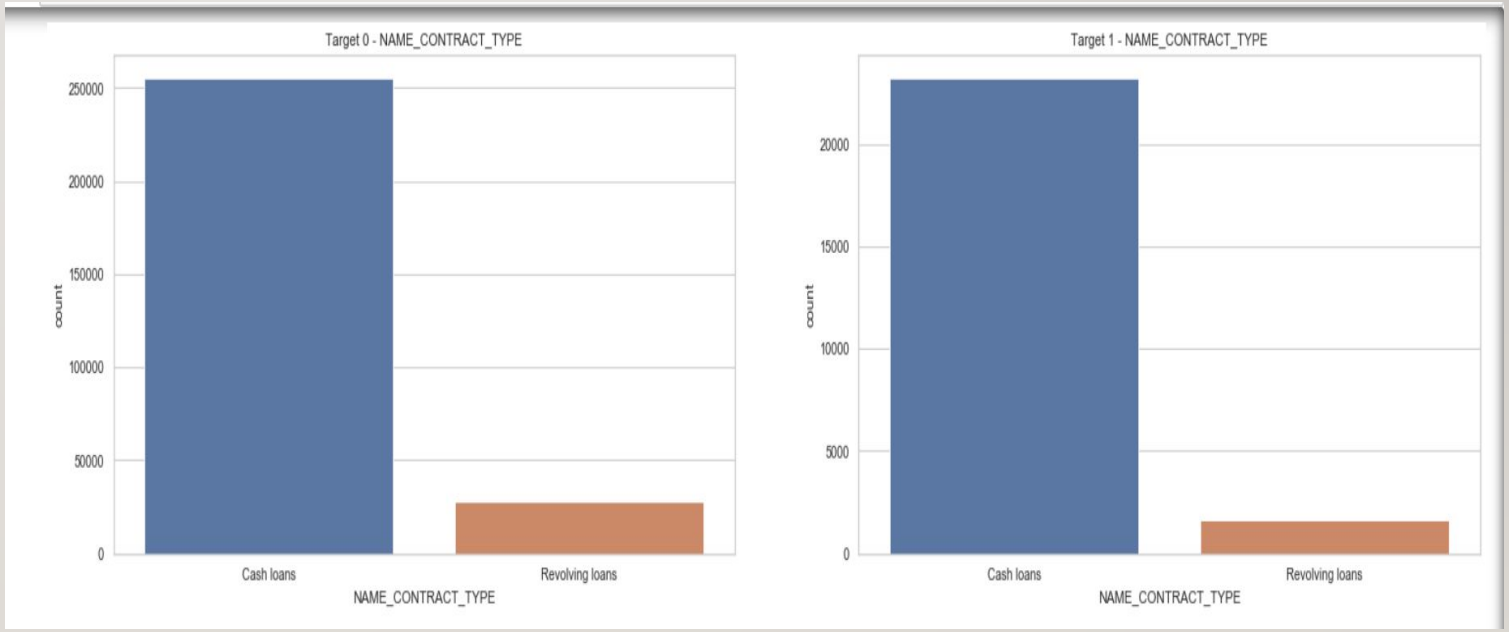
```
application_data.TARGET.value_counts(normalize=True)*100
```

```
0    91.927118
1     8.072882
Name: TARGET, dtype: float64
```

There is clearly imbalance in data. Person who is going to pay the loan EMI are very high(92%) and those who are going to default are less(8%)

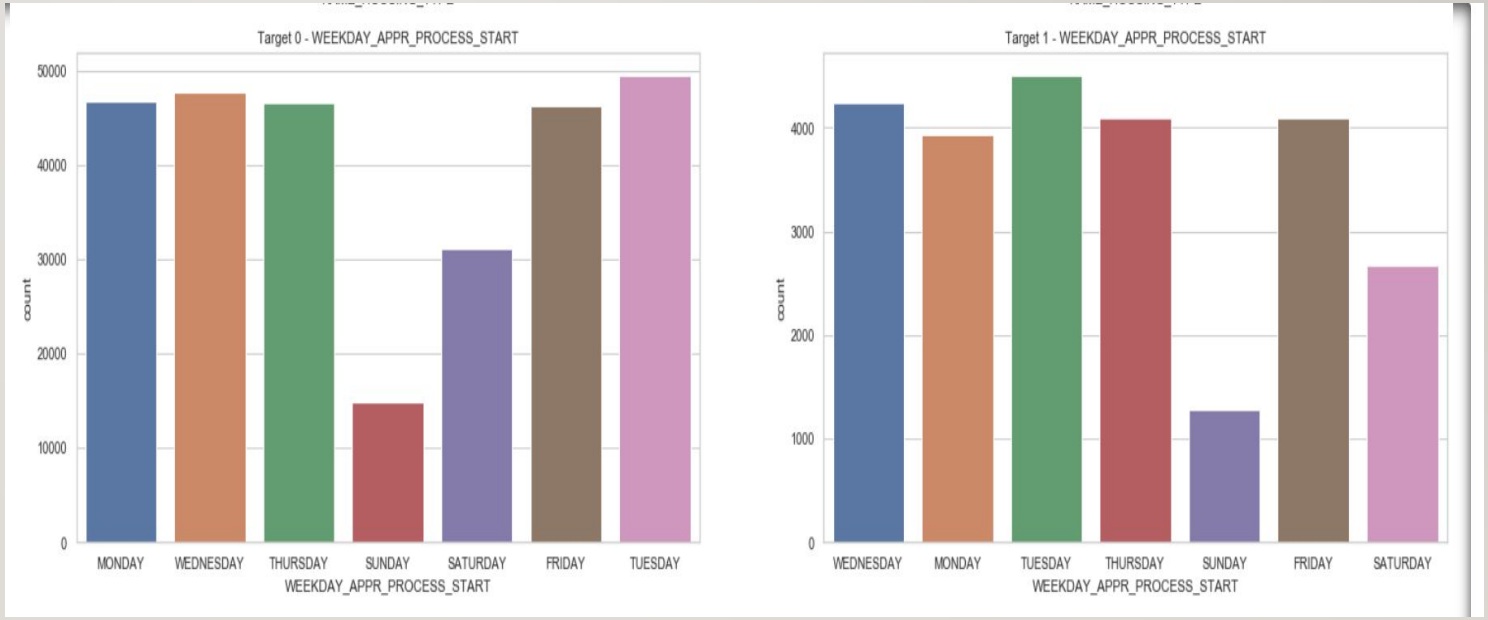
UNIVARIATE ANALYSIS: CATEGORICAL

- This graph shows the count of each contract type of both Target: 0 & 1.
- As per graph, cash loans application are much more in numbers than resolving loans.
- Also, count for target 0 is slightly higher than target 1.



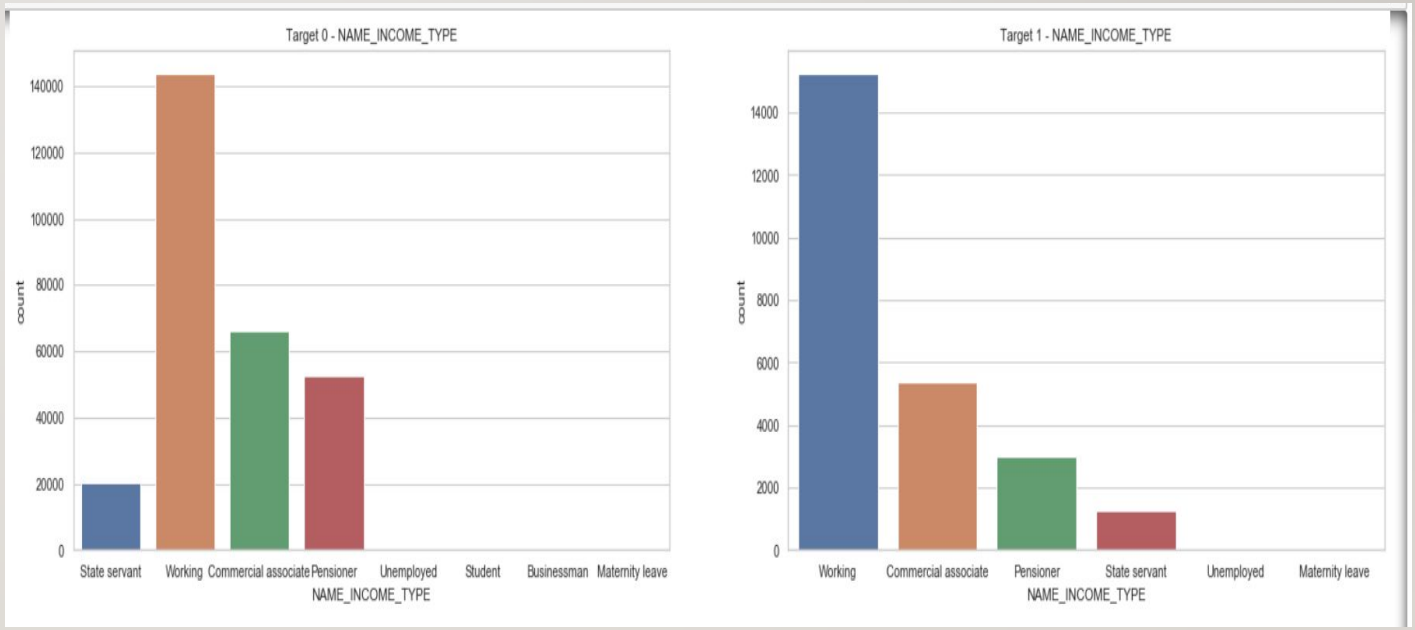
CONTINUE …..

- This graph shows the count of application days wise of both Target: 0 & 1.
- As per graph, for both target values, mostly application processed on Tuesday and least on Sunday .
- Here also, count for target 0 application is higher than target 1.



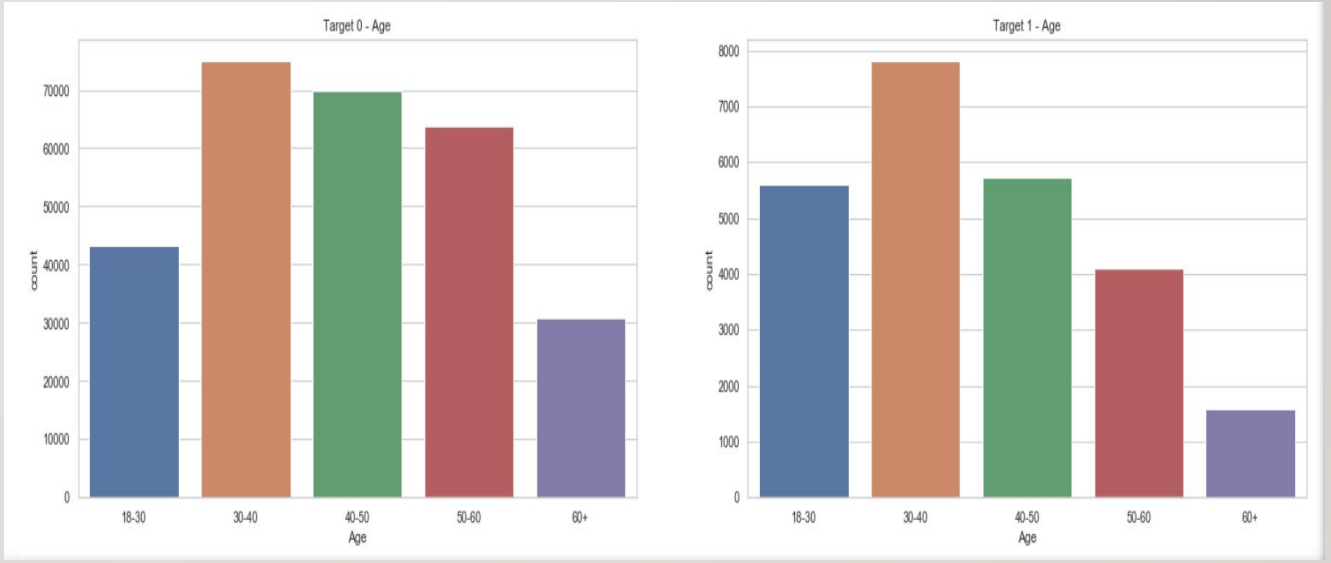
CONTINUE …..

- This graph shows the count of Target: 0 & 1 based on Income type.
- As per graph, for both target values, mostly application processed for Working type of people.
- Also, the category for target-0 are more than target1 like no Businessman & Student type application for target-1.



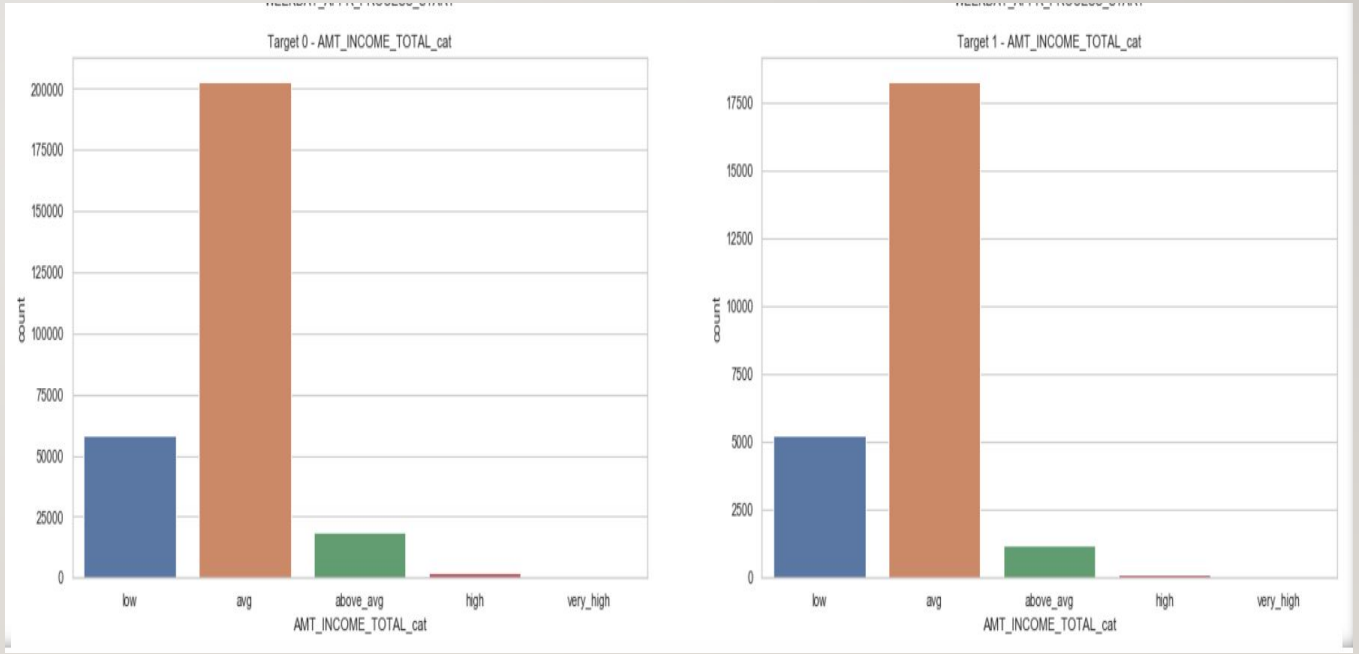
CONTINUE …..

- This graph shows the count of target: 0 & 1based on age binding column
- As per graph, mostly application for both target variable varies between 30-40 age group.
- For 18-30 age group data, target-1 applications are higher than target0.
- For 60+ age group data, target-1 applications are lesser than target0.
- For age group 30-40, 40-50 and 50-60, there is slight difference count for target 0 but for target1 difference is clearly visible.



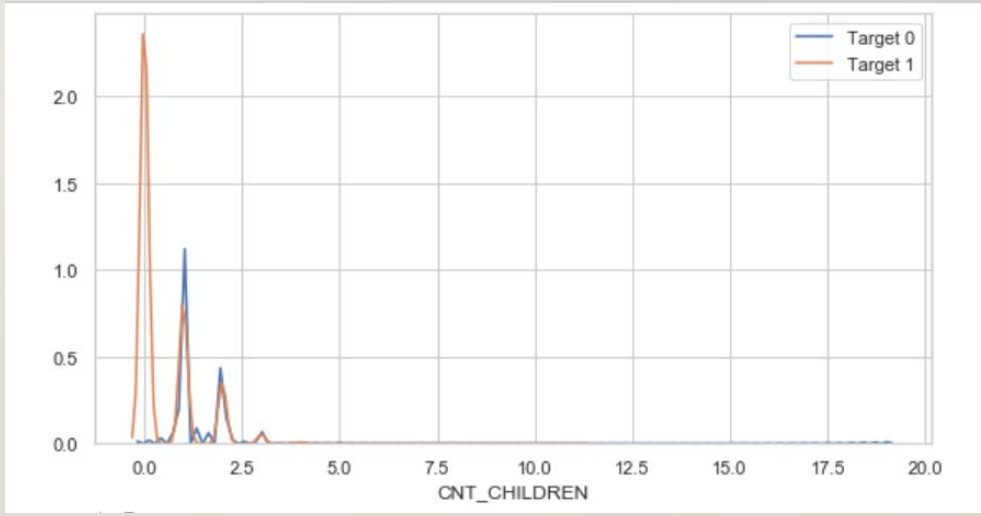
CONTINUE …..

- This graph shows the count of target: 0 & 1based on total income binding column.
- As per graph, mostly application for both target variable are from average salary group.
- For other groups, there is almost common pattern is visible for both target values.



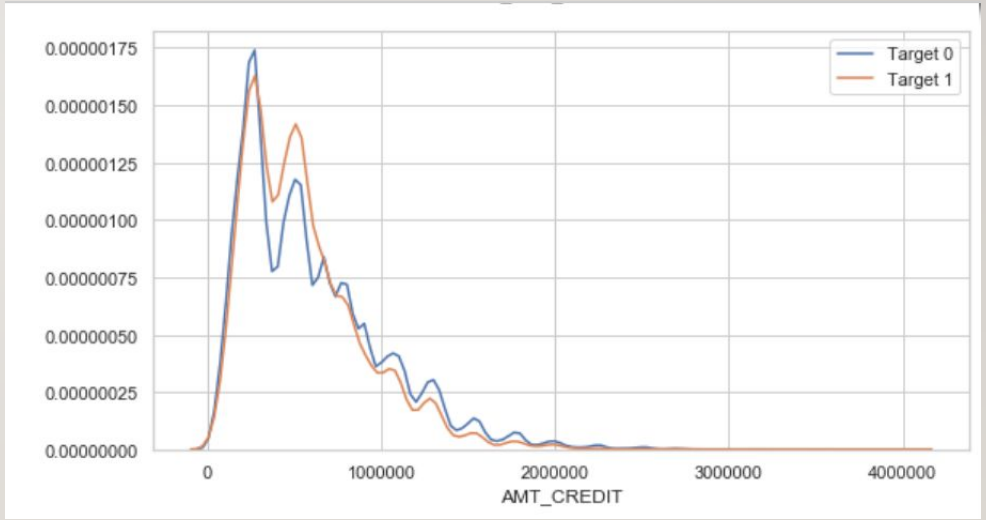
UNIVARIATE ANALYSIS: CONTINUOUS

- This graph shows the variation of target: 0 & 1based on children count category.
- As per graph, for target 0 the count is much higher than target 1 values in the starting range.
- As the range increase pattern doesn't show much variance for both values.
- In between range 0.0-2.5, target 1 values are higher than 0 but after that again more or less variance is same for both values.



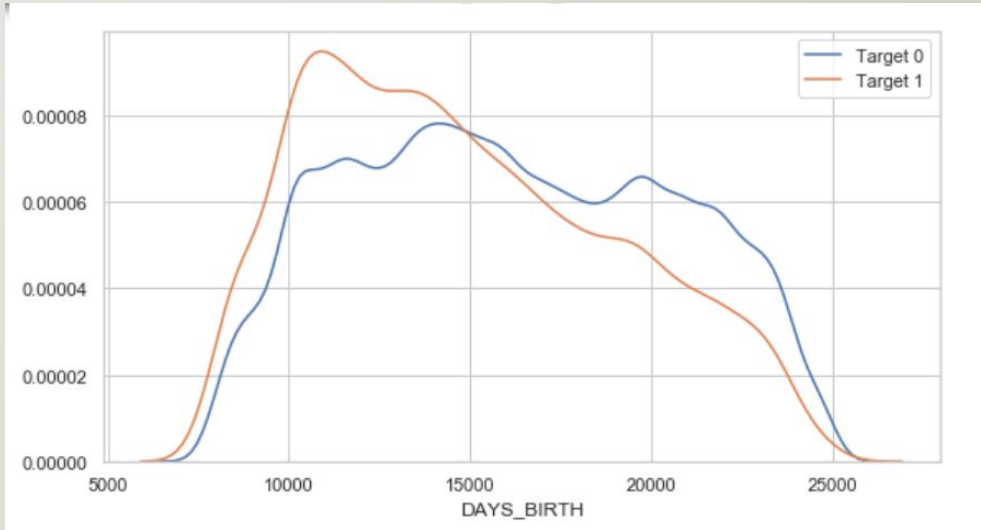
CONTINUE …..

- This graph shows the variation of target: 0 & 1based on amount credit category.
- As per graph, for target 0 the count is slightly higher than target 1 values in the range starting from 0.
- In between range 0-10Lc, for target 0 amount variance pattern is more than target 1. It means amount variant more for target 0 data.
- As the range increase pattern doesn't show much variance for both values.



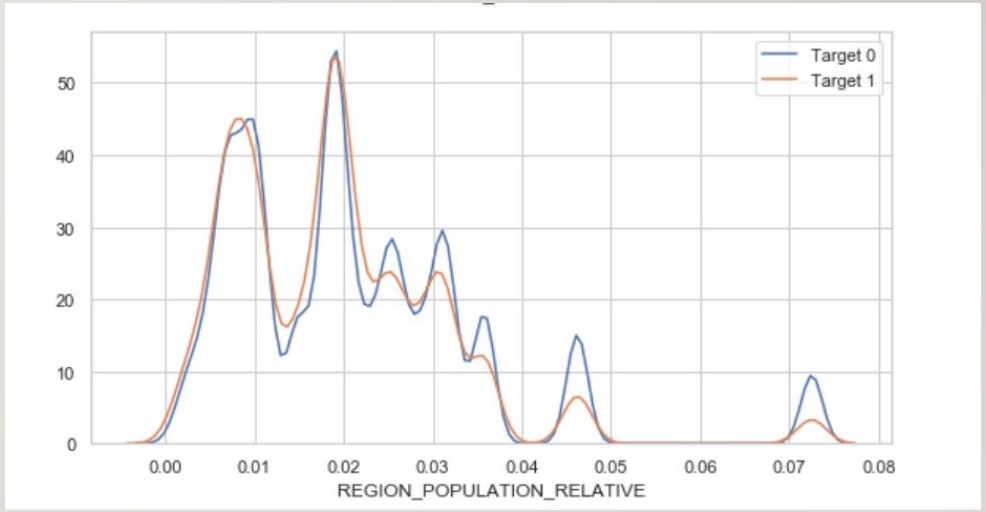
CONTINUE …..

- This graph shows the variation of target: 0 & 1based on client's age in days at the time of application.
- As per graph, for target 0: most application for ~11000 days (~28 year) and count decrease from 20000 days (~55 year) .
- As per graph, for target 1: most application for ~14000 days (~38 year) and count decrease from ~22000 days (~60 year) .
- We can say that more application of younger age are for target 0 whereas for senior citizens target 1 applications are higher.



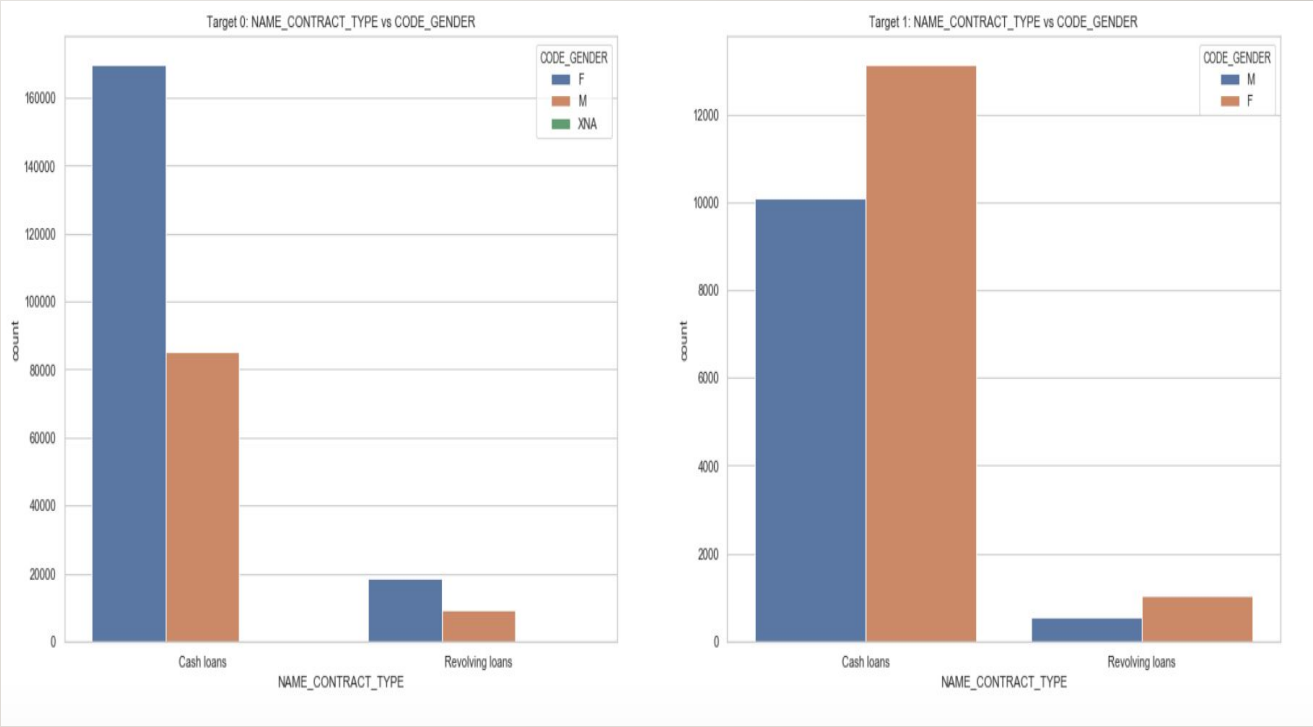
CONTINUE …..

- This graph shows the variation of target: 0 & 1based on normalized population of region where client lives (higher number means the client lives in more populated region).
- For both target values, graph variance pattern is nearly equal.
- As the range increase, count for target 0 is higher than 1.
- We can say that, for both target values normalized population of region division is same.



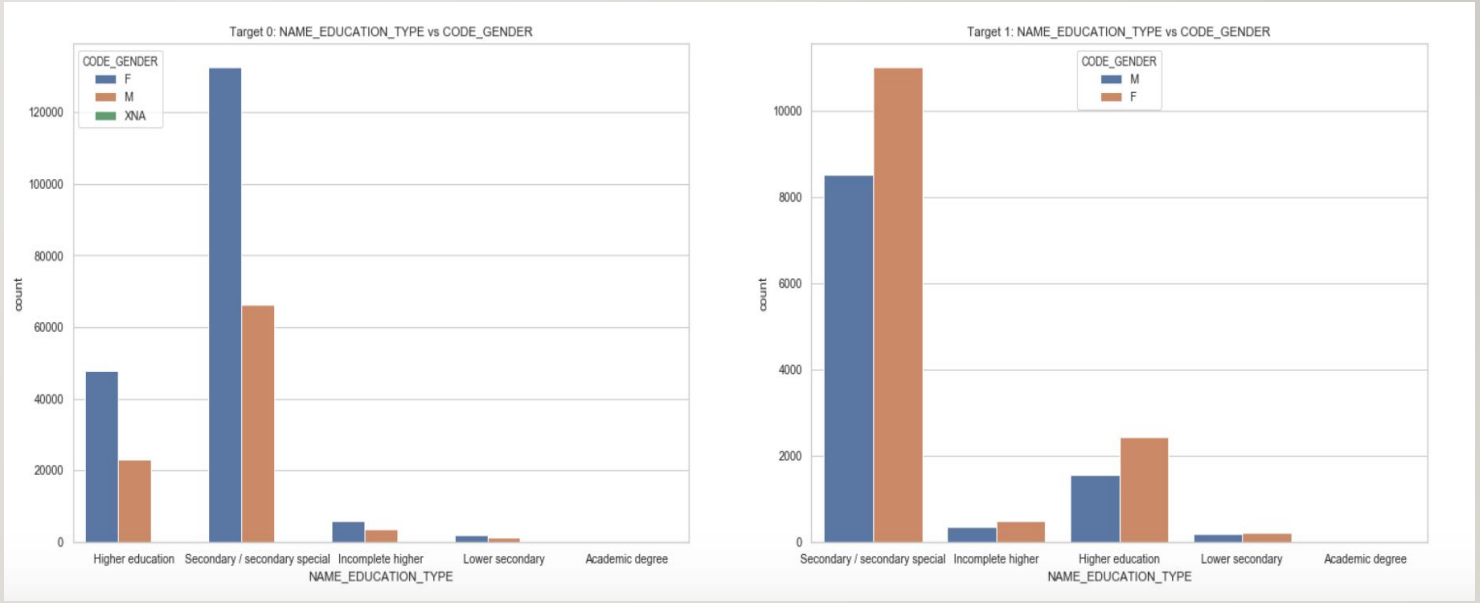
BIVARIATE ANALYSIS: CATEGORICAL - CATEGORICAL

- This graph shows the difference between loan Contract type and Gender for both target values.
- As per graph, it is clearly shown for both target values, Female gender applied loan application than male.
- Also, count difference for Female and Male gender applications is higher for Target0- Cash Loan types.
- For resolving loans type count is lower but Female are leading here as well.



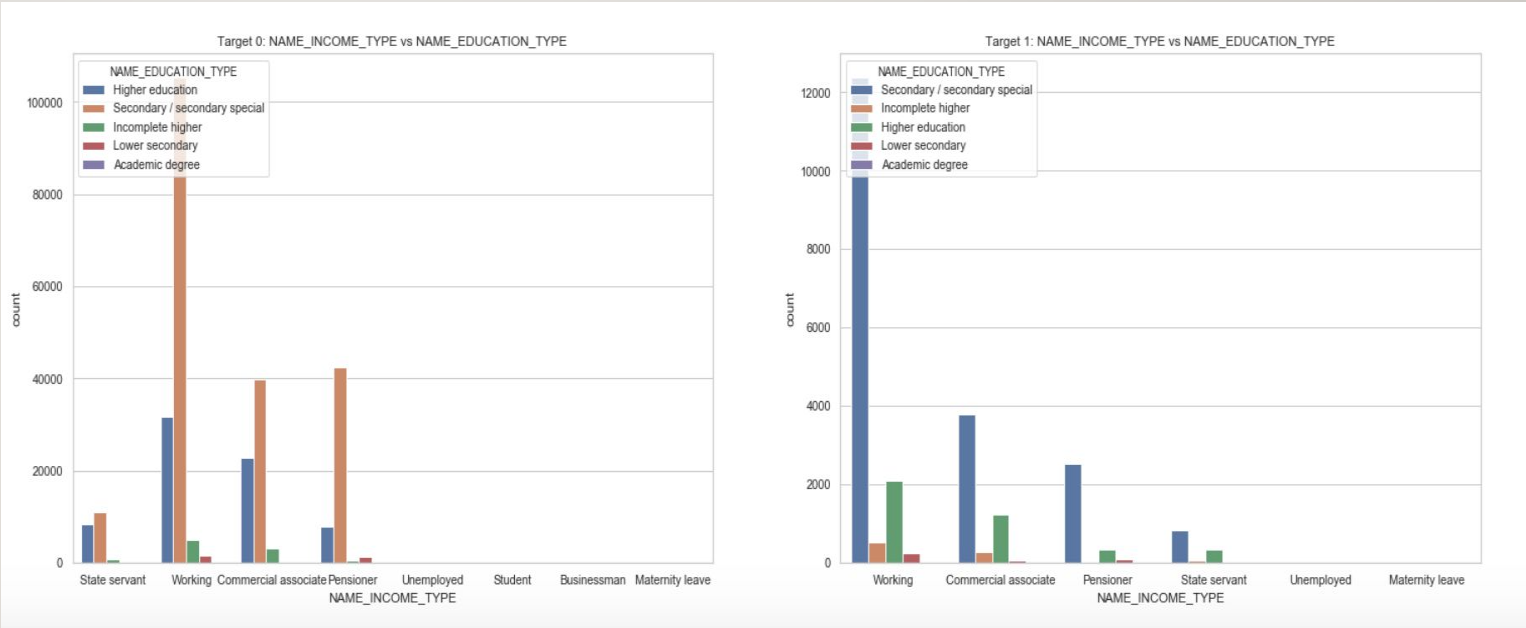
CONTINUE …..

- This graph shows the difference between loan Education type and Gender for both target values.
- As per graph, it is clearly shown for both target values, people who have Secondary/senior secondary education applied more loan application than all other categories.
- Here, also female application count is higher than male.
- The difference count for male and female is higher for target 0 values.



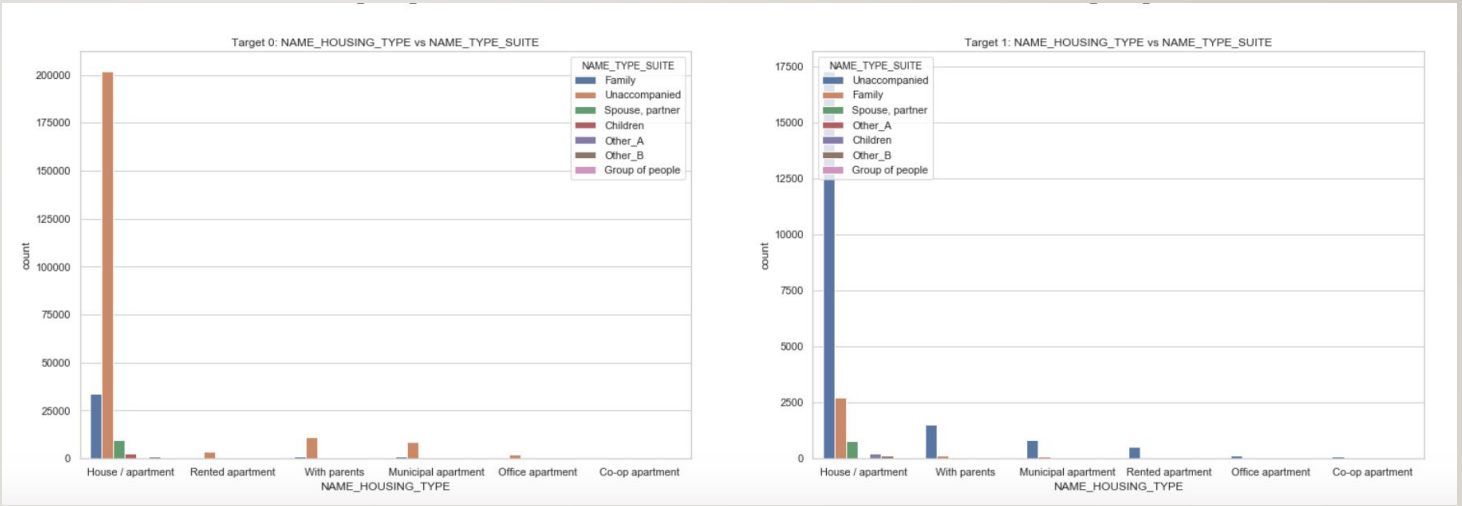
CONTINUE …..

- This graph shows the difference between loan Education type and Income type for both target values.
- As per graph, it is clearly shown for both target values, who lies in working income and Secondary education are highest.
- There is very smaller number of application for unemployed, businessman, maternity leave or student category.
- For target value 0 count is nearly equal for secondary education of commercial associate and pensioner whereas for target1 there is good difference.



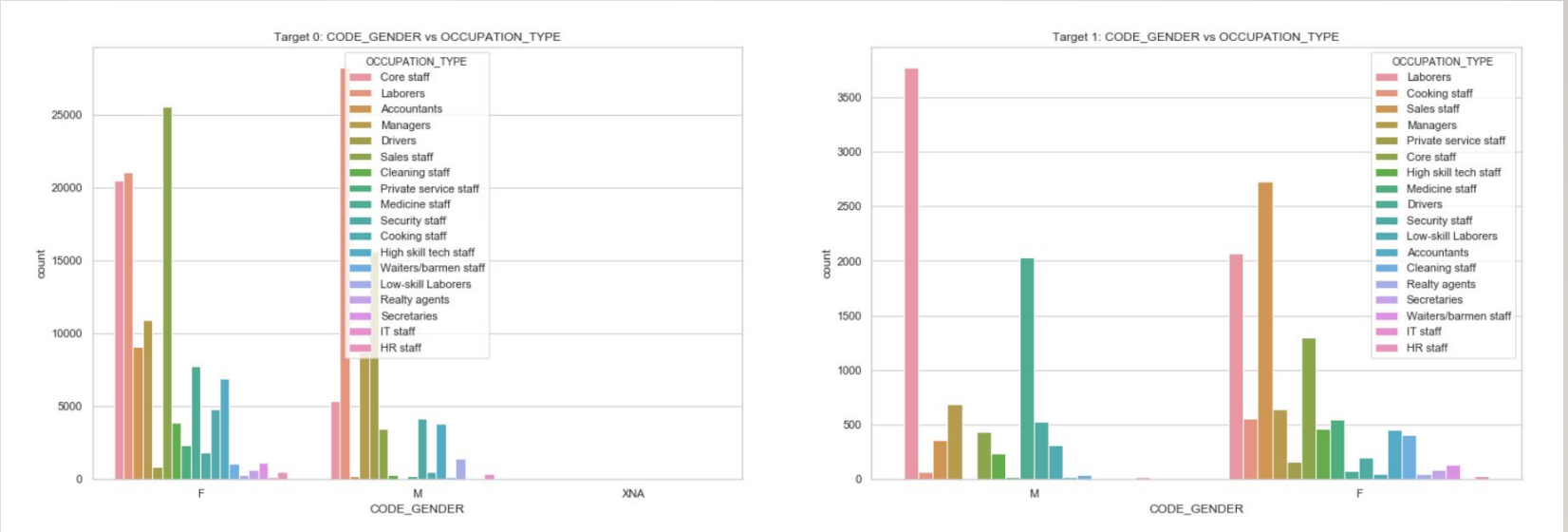
CONTINUE …..

- This graph shows the difference between loan Housing type and Family type for both target values.
- As per graph, it is clearly shown for both target values, people who have their own house/apartment and unaccompanied are highest in number.
- For other categories, count is very less and nearly equal for both target values.



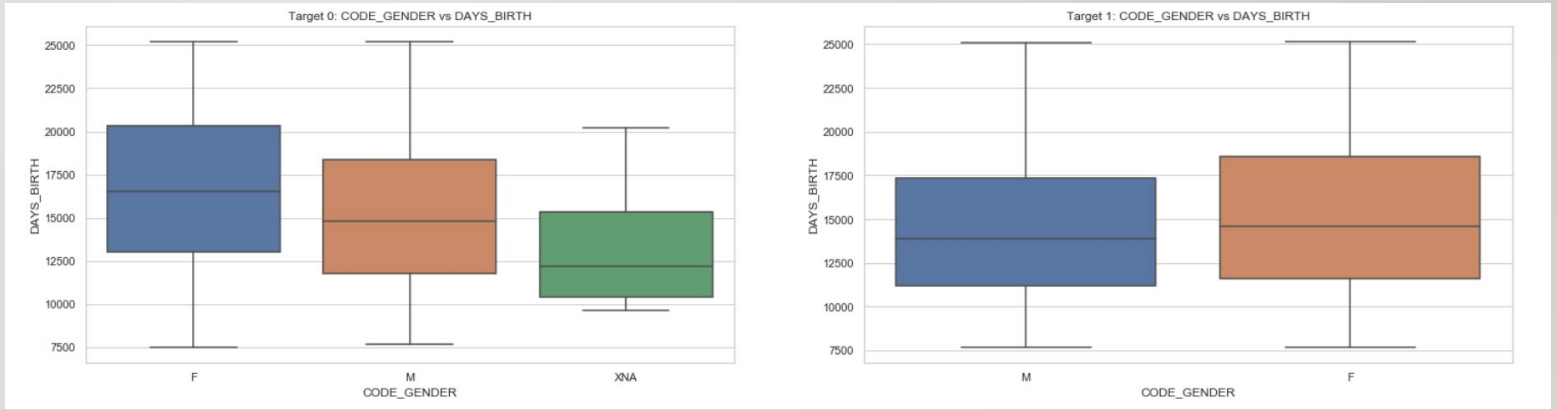
CONTINUE …..

- This graph shows the difference between loan Occupation type and Gender for both target values.
- As per graph, Males who have laborers occupation have highest application.
- For target1 application count is higher than target 0 for each occupation type.
- Total Variance patter for both target values is same.



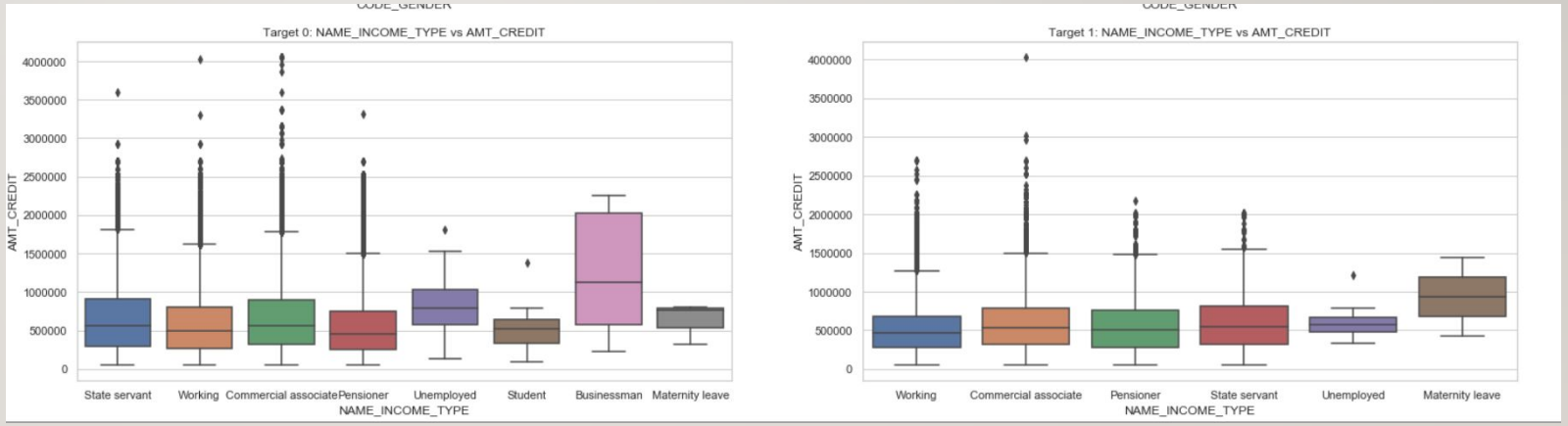
BIVARIATE ANALYSIS: CATEGORICAL - CONTINUOUS

- For both male and female minimum line is same of each target value.
- Median value of female target 0 is highest.
- As per graph, data imbalance for target 0 is clearly seen as null records are there.



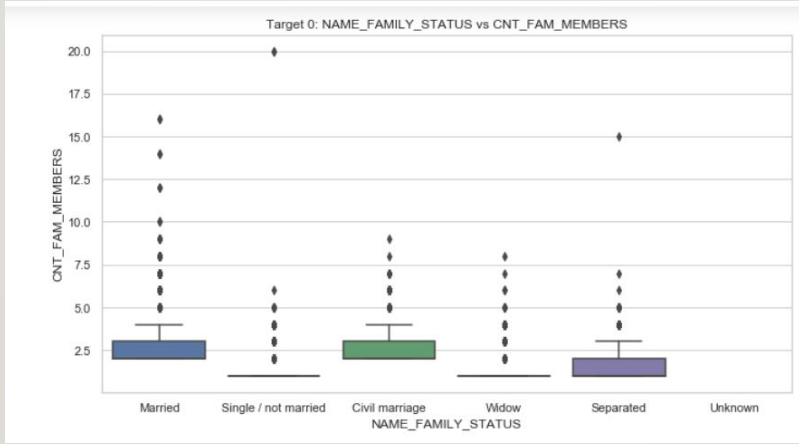
CONTINUE …..

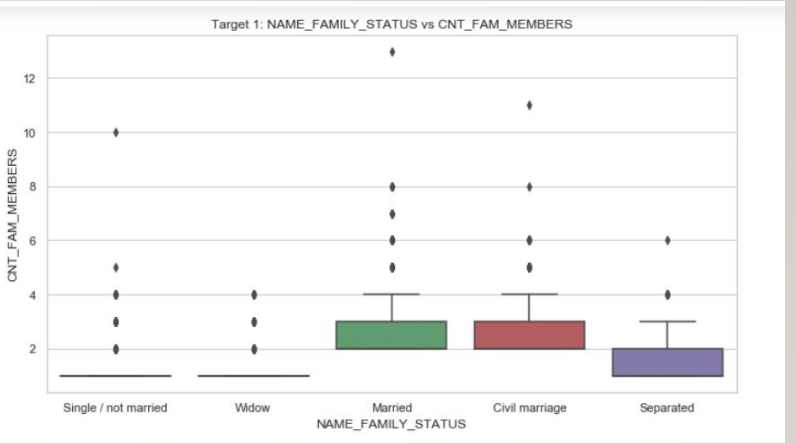
- Highest median for Target 0 value is for Businessman income type people.
- Highest median for Target 1 value is for Maternity leave income type people
- There are more outliers for for commercial associate category for both target values.
- Number of categorical classification is high for target 0 value.



CONTINUE …..

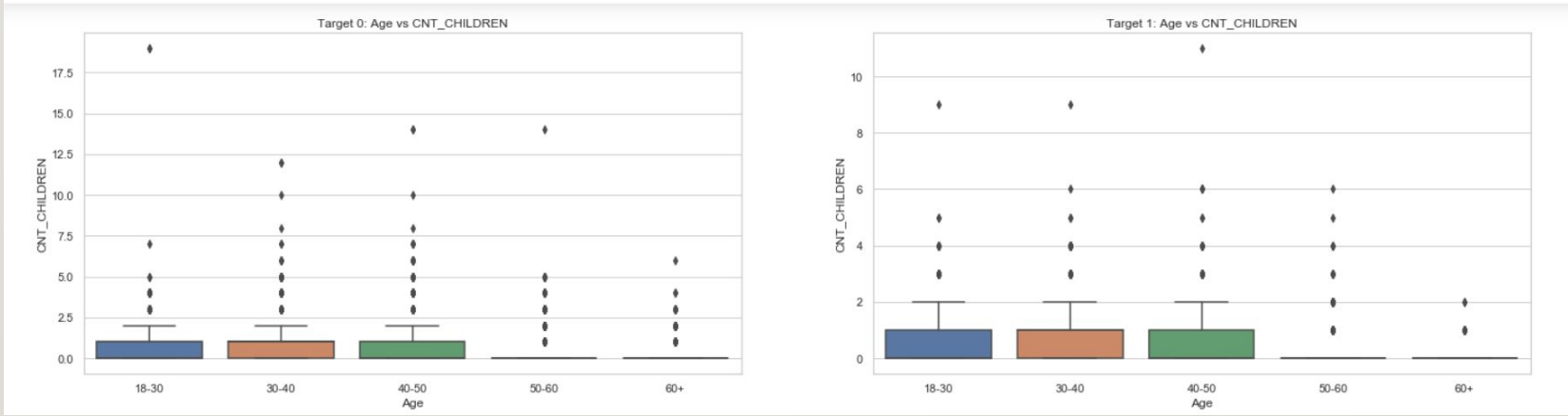
- Outliers for Married category is highest for both target values.
- For target 1 maximum number of married and civil marriage is same.
- For single and window category the valid family count member is zero. Also, having few outlier count for both target values.





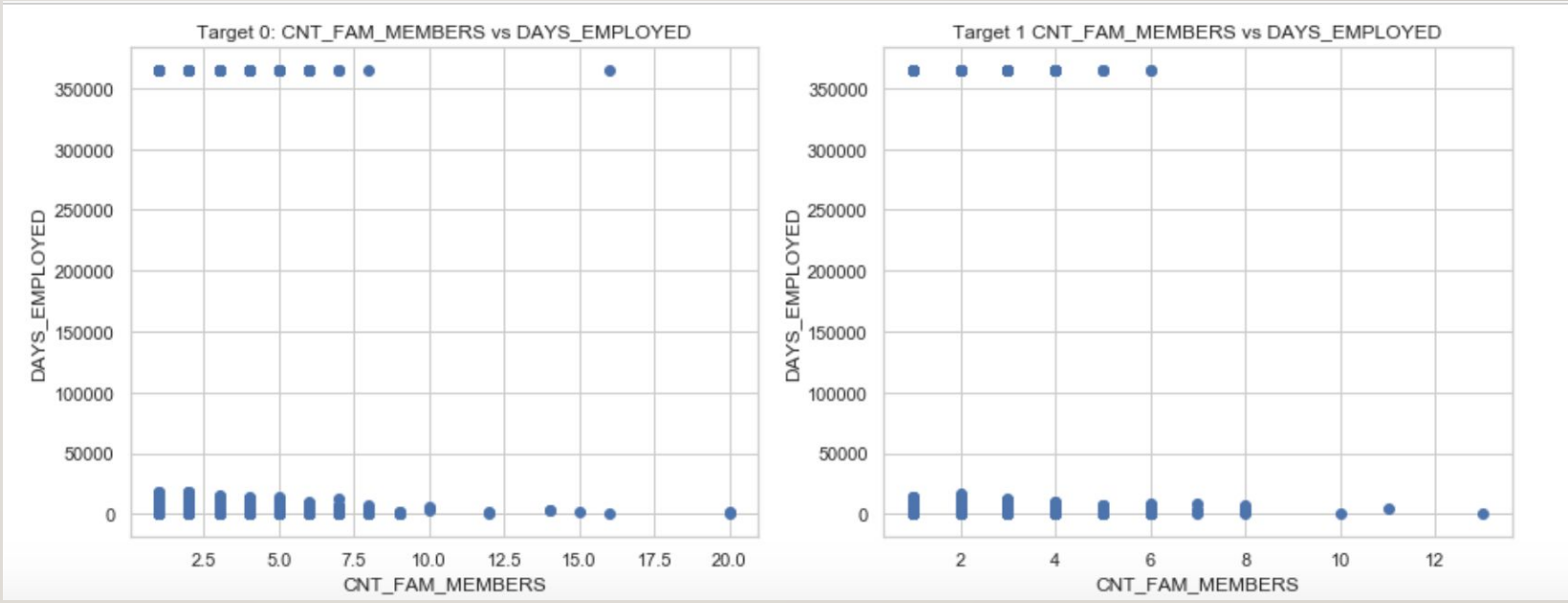
CONTINUE …..

- Maximum range for each age group of each target value is same.
- The highest outliers values is for 40-50 age group of target 0 value.
- The target 1 data is equally distributed for category 18-30, 30-40 and 40-50.



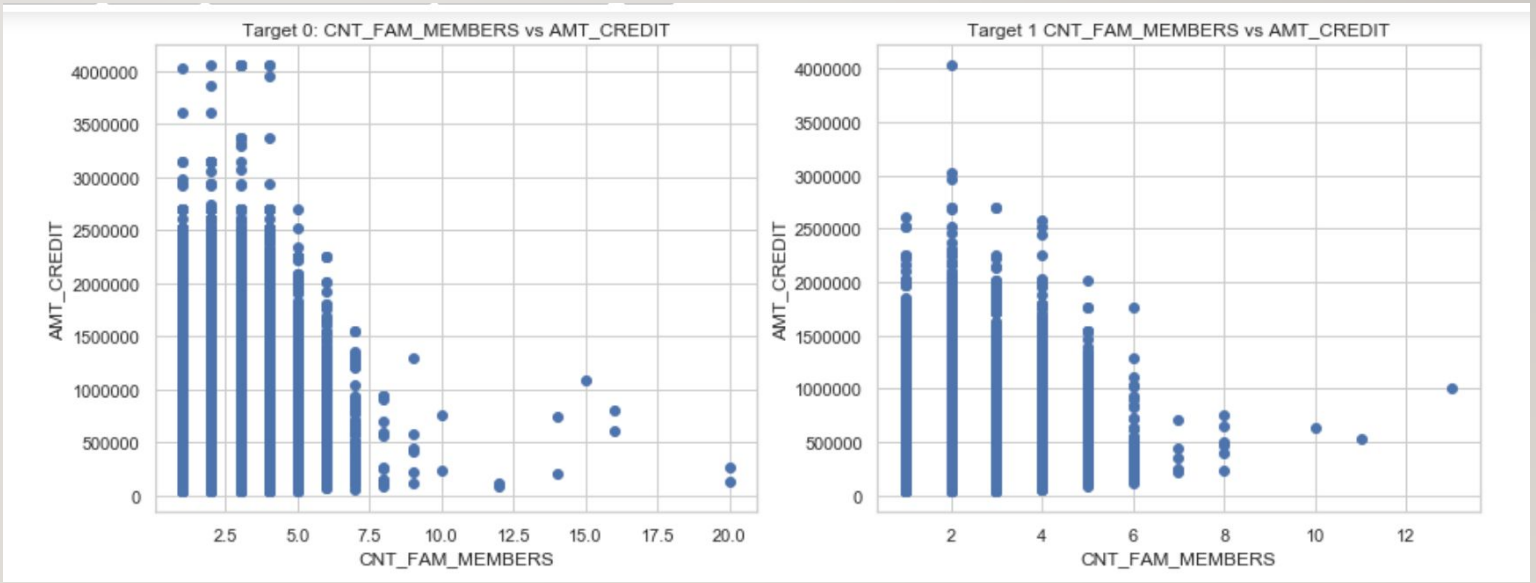
BIVARIATE ANALYSIS: CONTINUOUS - CONTINUOUS

- This graph shows the difference between family member count and how many days before the application the person started current employment.
- The variance pattern is almost same for both target values.
- For family member range 15-17.5 and highest range of employed days we have few application for target 0 but not for target 1.
- As per graph days employed range goes to 350K days(~958 years) which is impossible range. It means either we don't have proper data for this and consider as default value or incorrect data has been entered.



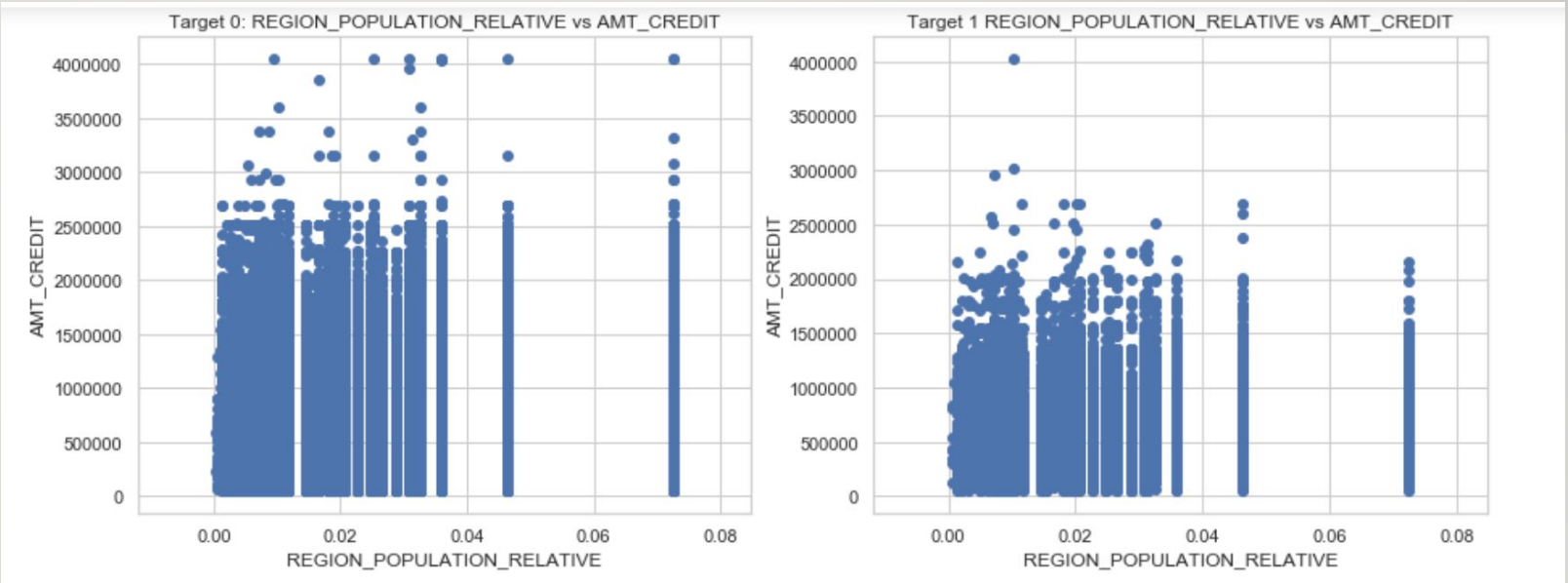
CONTINUE …..

- This graph shows the difference between family member count and loan amount credit
- The variance pattern is almost same for both target values and the difference is due to count of applications.
- As per graph the family member count range for target 0 varies from 0-20+ whereas for target 1 it is 0-12+.
- The count for highest credit amount ie. 40L is higher for target 0 value.
- Mostly application are from family member range of 0-5 and loan amount from 5-25 Lc.



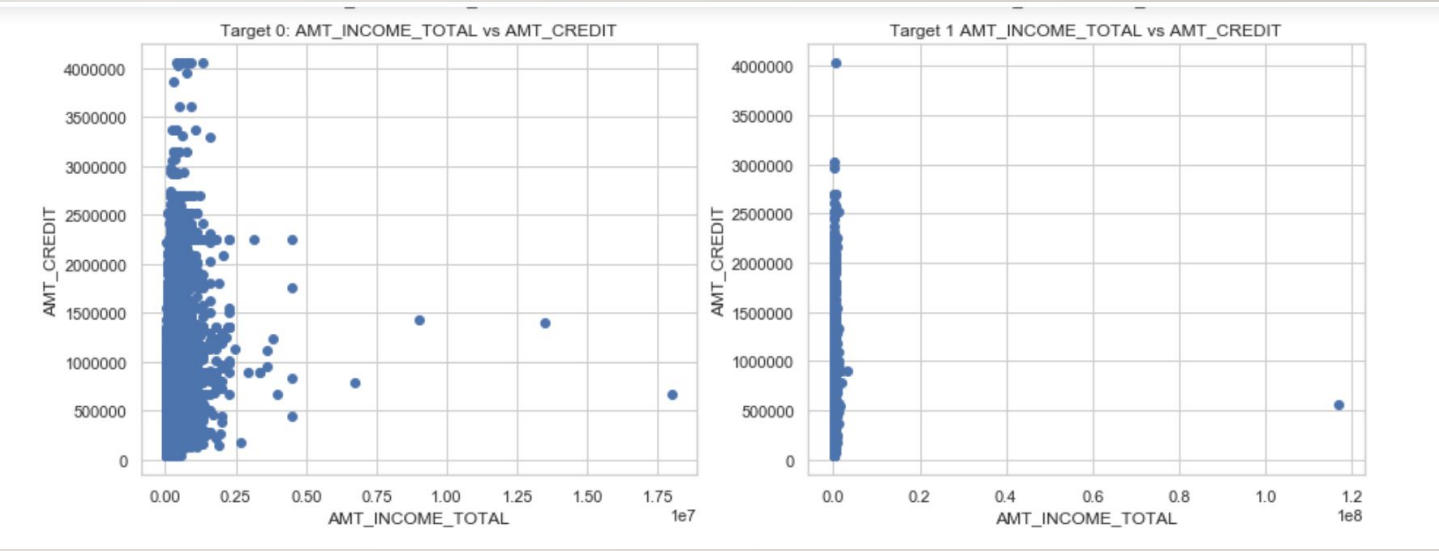
CONTINUE …..

- This graph shows the difference between region population and amount credit.
- The variance pattern is almost same for both target values and the difference is due to count of applications.
- As per graph, people who lives in more populated area have lesser application than least population area.
- Mostly application for 0.00-0.04 population range and almost same loan credit amount has been seen.



CONTINUE …..

- This graph shows the difference between person total income and loan amount.
- The variance pattern for target 0 is more than target 1 due to the application count difference.
- There are very few applications who have loan credit amount range from 30-40Lc for target 1.
- Similarly, for target 0 total income variance is higher than target 1.



CORRELATION – TARGET 0

- Top 3 correlation for target 0 are –
  - OBS_60_CNT_SOCIAL_CIRCLE with OBS_30_CNT_SOCIAL_CIRCLE
  - FLOORSMAX_MEDI with FLOORSMAX_AVG
  - YEARS_BEGINEXPLUATATION_MEDI with YEARS_BEGINEXPLUATATION_AVG

	Var1	Var2	Correlation	Correlation_Abs
666	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508	0.998508
576	FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997018	0.997018
547	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.993582	0.993582
578	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.988153	0.988153
143	AMT_GOODS_PRICE	AMT_CREDIT	0.987250	0.987250
520	FLOORSMAX_MODE	FLOORSMAX_AVG	0.985603	0.985603
491	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.971032	0.971032
549	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.962064	0.962064
309	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571	0.878571
144	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686	0.776686

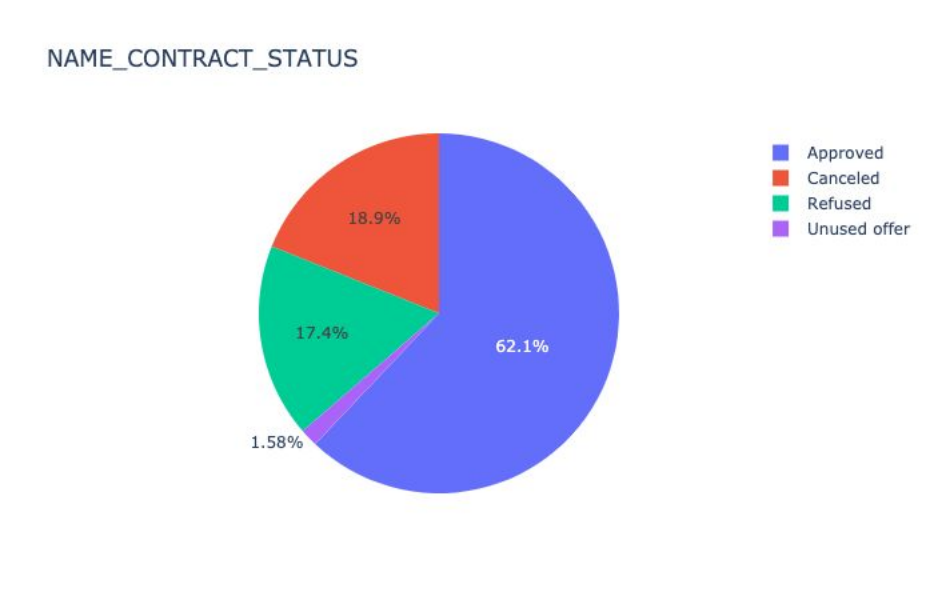
CORRELATION – TARGET 1

- Top 3 correlation for target 1 are –
  - OBS_60_CNT_SOCIAL_CIRCLE with OBS_30_CNT_SOCIAL_CIRCLE
  - FLOORSMAX_MEDI with FLOORSMAX_AVG
  - YEARS_BEGINEXPLUATATION_MEDI with YEARS_BEGINEXPLUATATION_AVG

	Var1	Var2	Correlation	Correlation_Abs
666	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269	0.998269
576	FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997187	0.997187
547	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.996124	0.996124
578	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.989195	0.989195
520	FLOORSMAX_MODE	FLOORSMAX_AVG	0.986594	0.986594
143	AMT_GOODS_PRICE	AMT_CREDIT	0.983103	0.983103
491	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.980466	0.980466
549	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.978073	0.978073
309	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484	0.885484
144	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699	0.752699

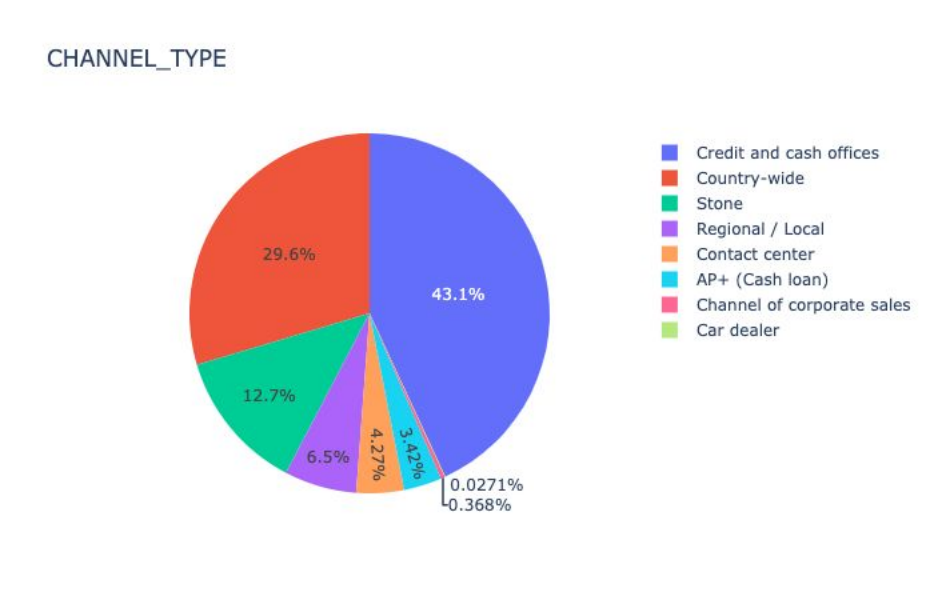
PREVIOUS DATA – UNIVARIATE ANALYSIS: CATEGORICAL

- As per graph these are the following division -
  - Approved: 62.1% times
  - Cancelled: 18.9 % times
  - Refused: 17.4 % times
  - Unused offer: 1.58 % times
- Which means more than half of applications are approved successfully.



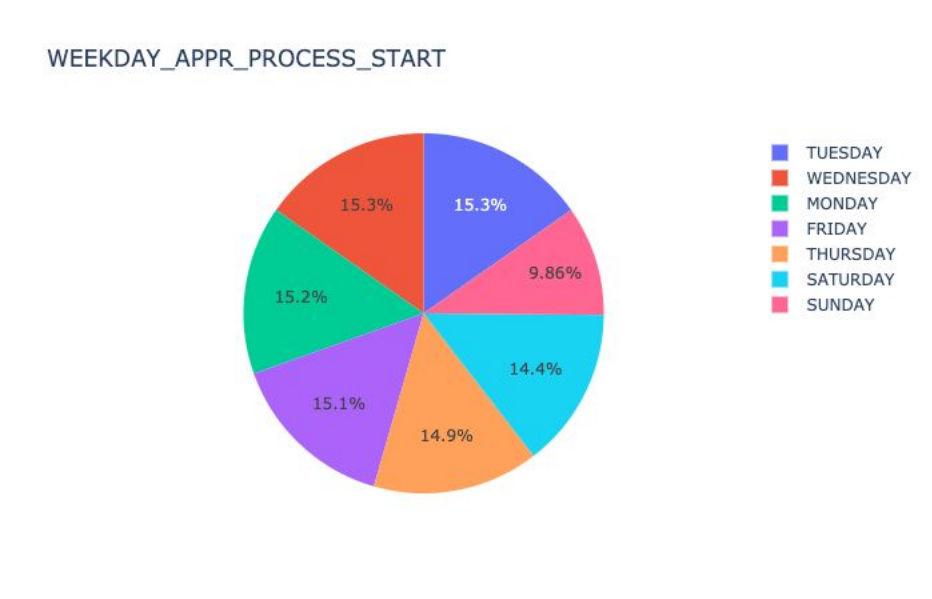
CONTINUE …..

- 43% applications were filed through credit & cash offices, which is highest.
- 29.6% were filed through country-wide.
- Other channel type consists of remaining channel type which is around less than equal to 12%.



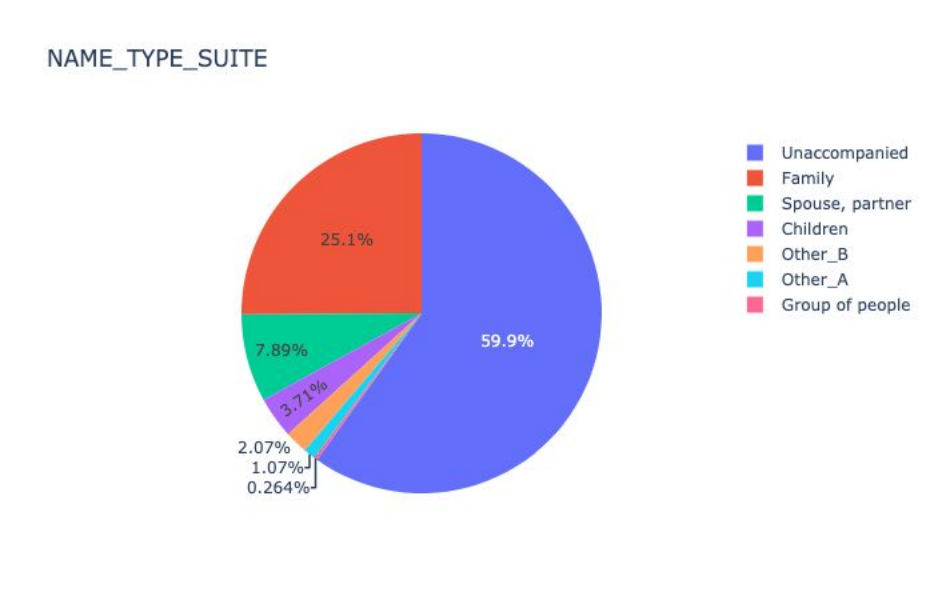
CONTINUE …..

- Most of the application were processed on weekdays.
- As per graph, ~25% of total application processed on weekend.
- Application processing percentage of any day on weekday is almost equal.



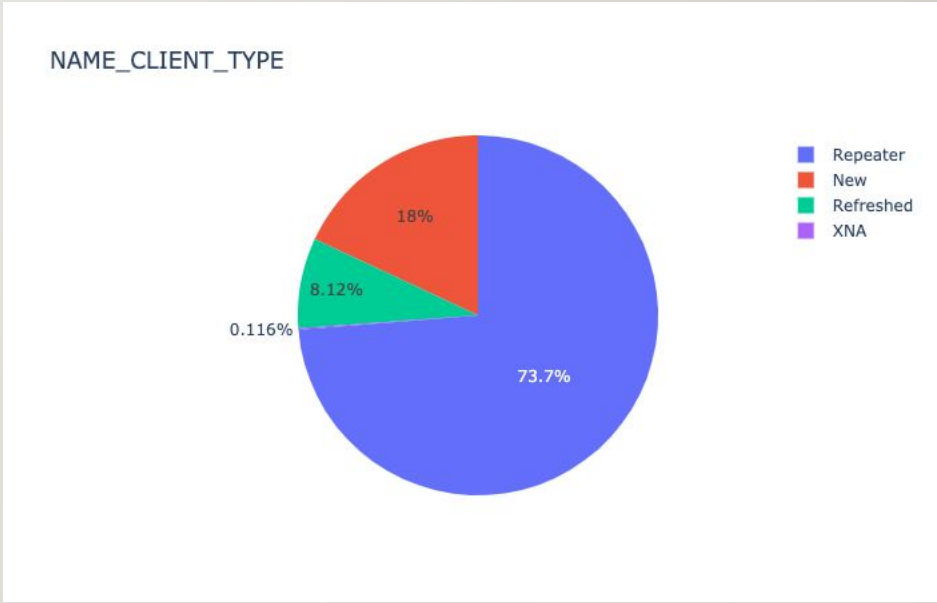
CONTINUE …..

- 60% people were unaccompanied while filing the applications.
- 25% were with family and 8% were with spouse.



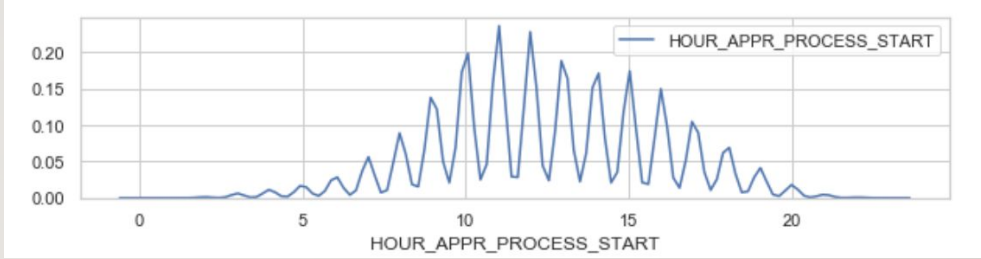
CONTINUE …..

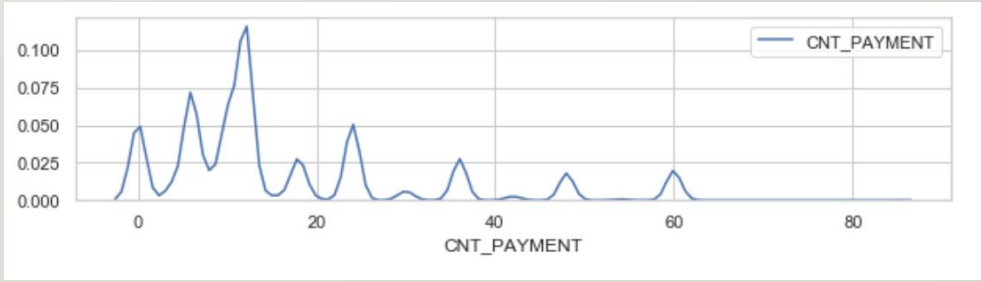
- 77% previous applications were repeated.
- While only 18% were new applications.



PREVIOUS DATA – UNIVARIATE ANALYSIS: CONTINUOUS

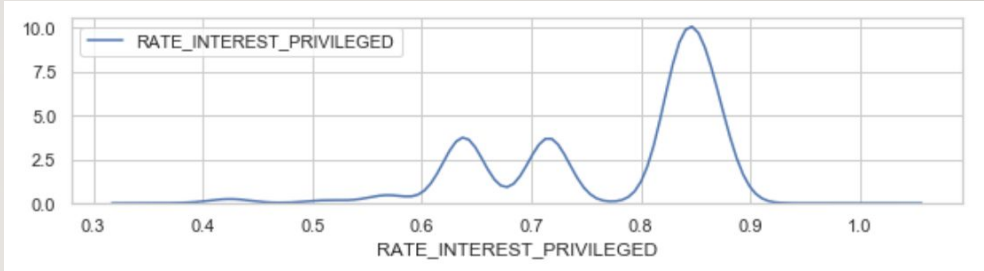
- **Hour_Appr_Process_Start:** Most of the applications were processed between 5-15 hours.
- **CNT_Payment:** It varies between 0 to 80 but maximum application had between 0 to 20.

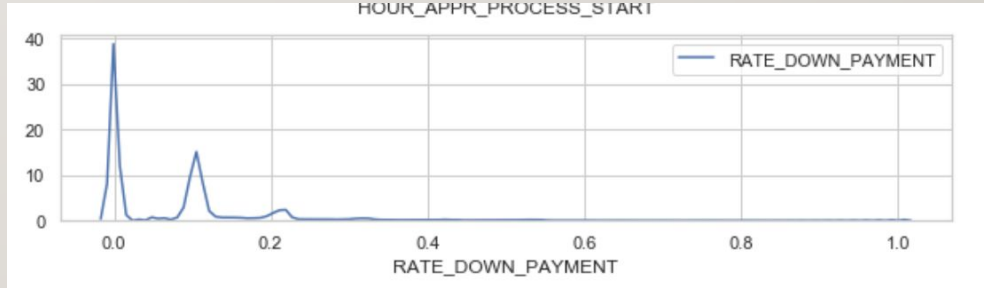




CONTINUE …..

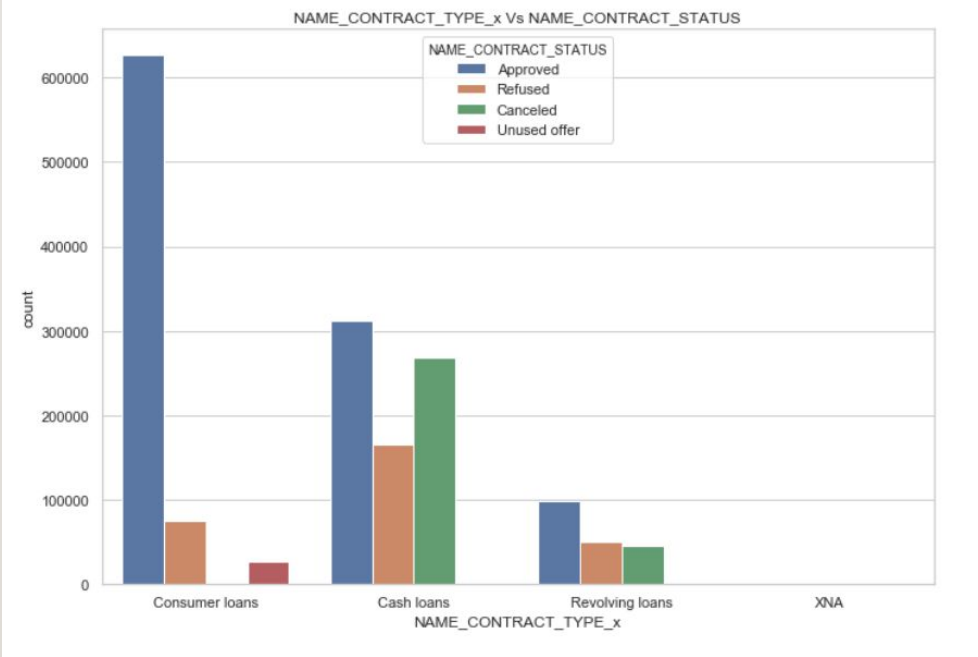
- **Rate_Interest_Privileged:** Interest rate varies between 0.3 to 1 but 0.8 to 0.9 had the highest.
- **Rate_Down-Payment:** Down payment rate varies between 0 to 1 but majority had between 0 to 0.1





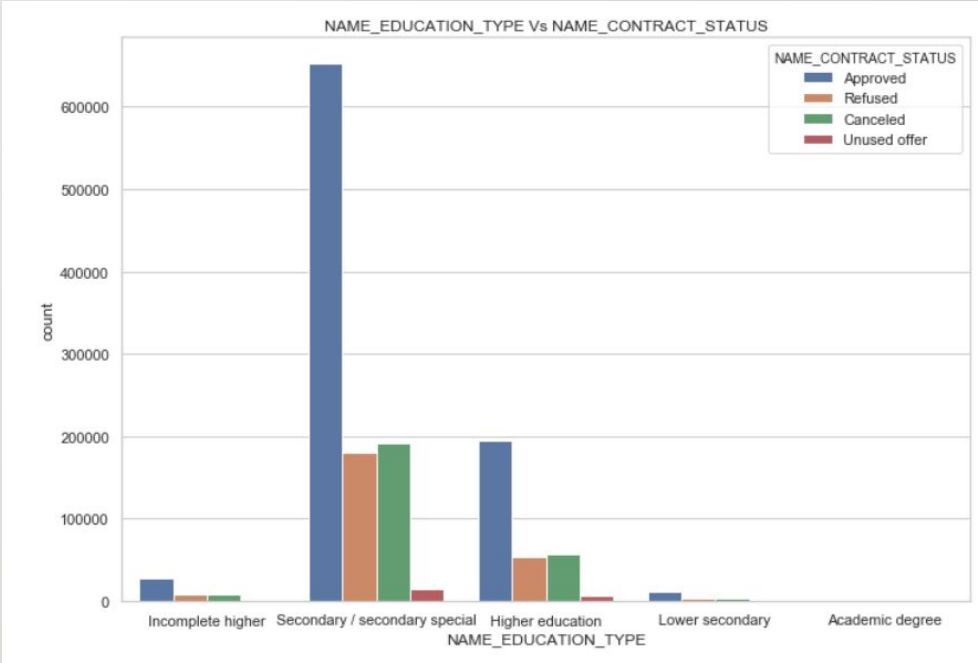
PREVIOUS + APPLICATION DATA – BIVARIATE ANALYSIS: CATEGORICAL - CATEGORICAL

- After merging both file data we see one new category Consumer loans is added which is not present in application data.
- Consumers loans were mainly approved.
- Mostly cash loans are refused or cancelled in comparative of consumer loans.
- Half of the Revolving loans were approved while half were cancelled or refused.



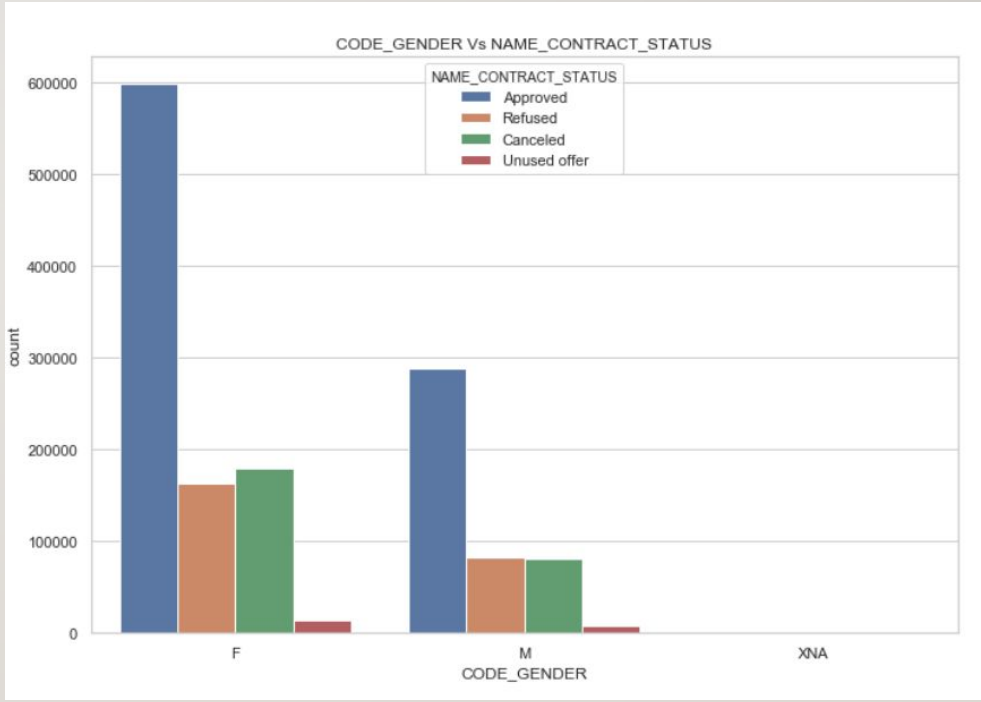
CONTINUE …..

- For secondary/secondary special data we have Highest approval as well as cancelled applications.
- While count for unused offer is least or negligible for all education type.



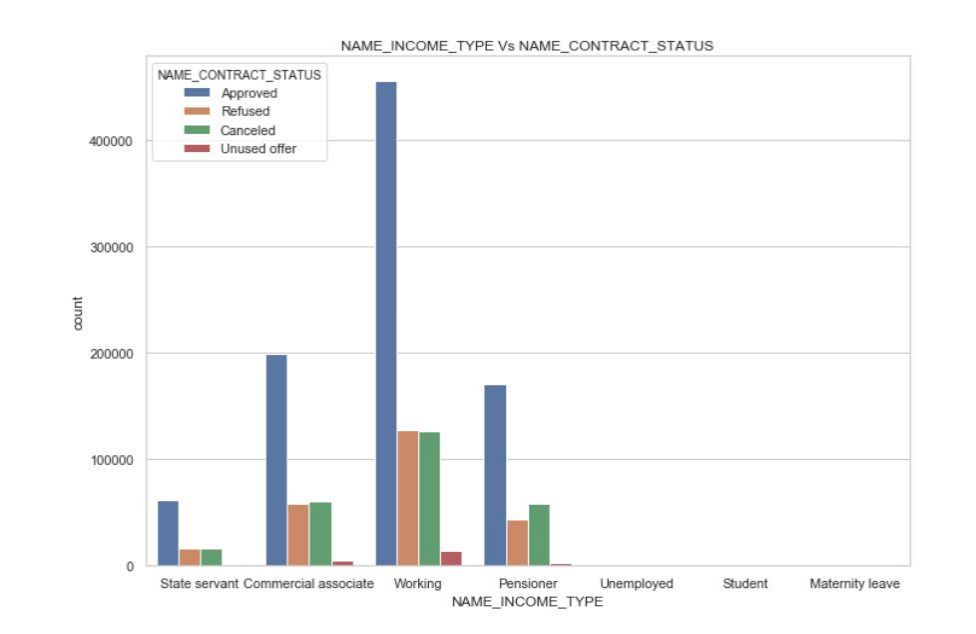
CONTINUE …..

- The ratio of female loan approved and cancelled is higher than male.
- There is almost same ratio for refused and cancelled application for male gender.



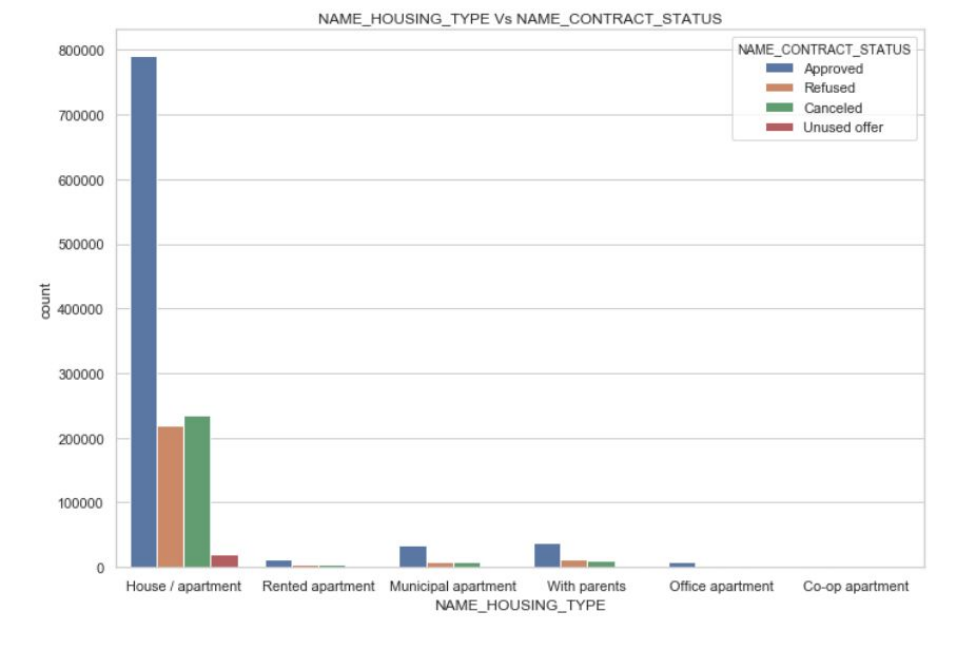
CONTINUE …..

- Working people got most approval of their loan.
- Pensioners loan having most cancelled count.
- For unemployed, student and maternity leave count value is negligible.



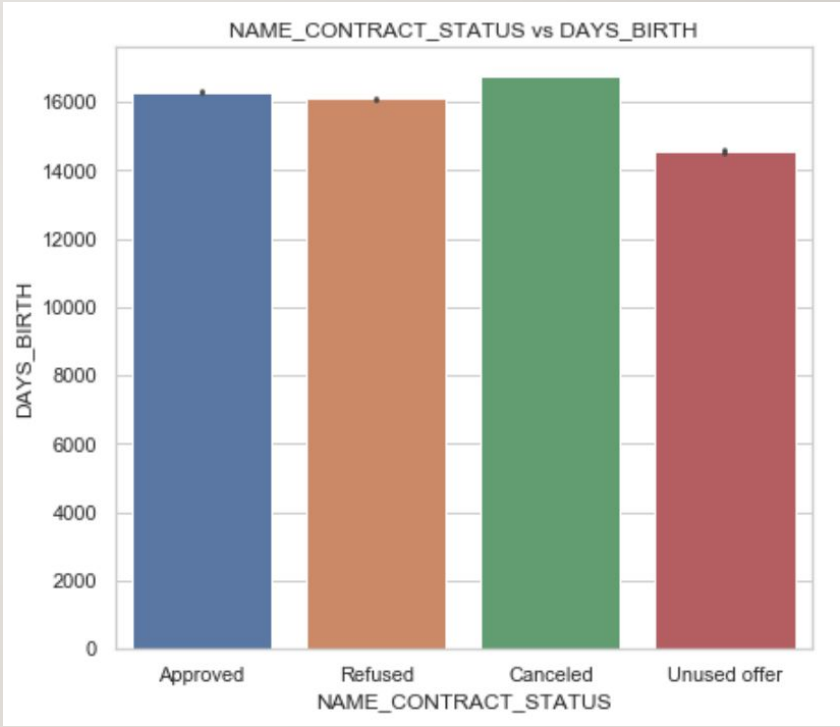
CONTINUE …..

- People living in home/apartment got most approval as well as cancellation and refusal.
- For others we have less amount of data is present.



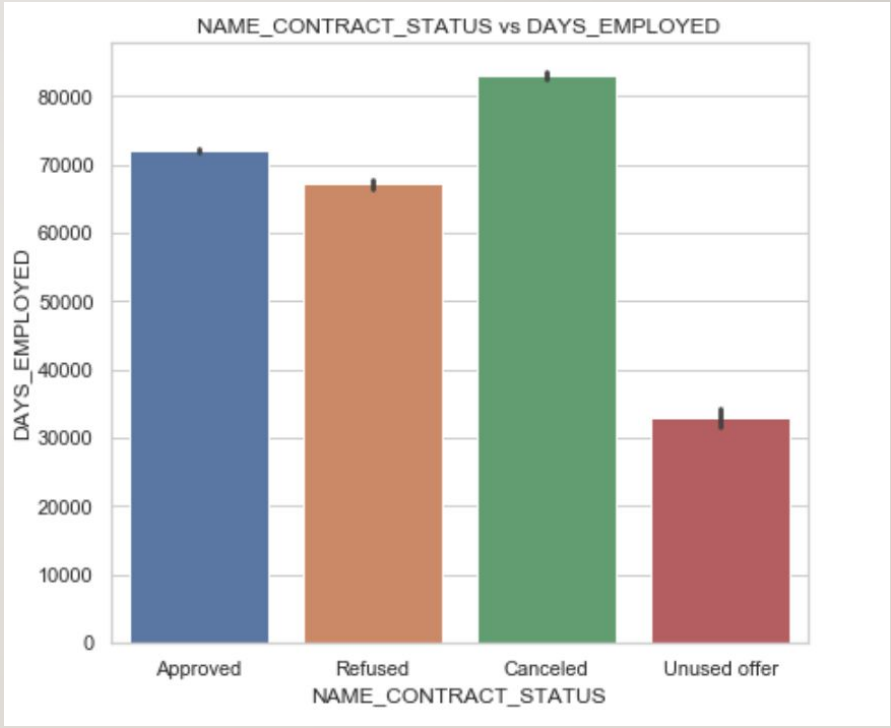
PREVIOUS + APPLICATION DATA – BIVARIATE ANALYSIS: CATEGORICAL - CONTINUOUS

- Count for cancelled application is highest.
- There is slight difference for approved and refused application.



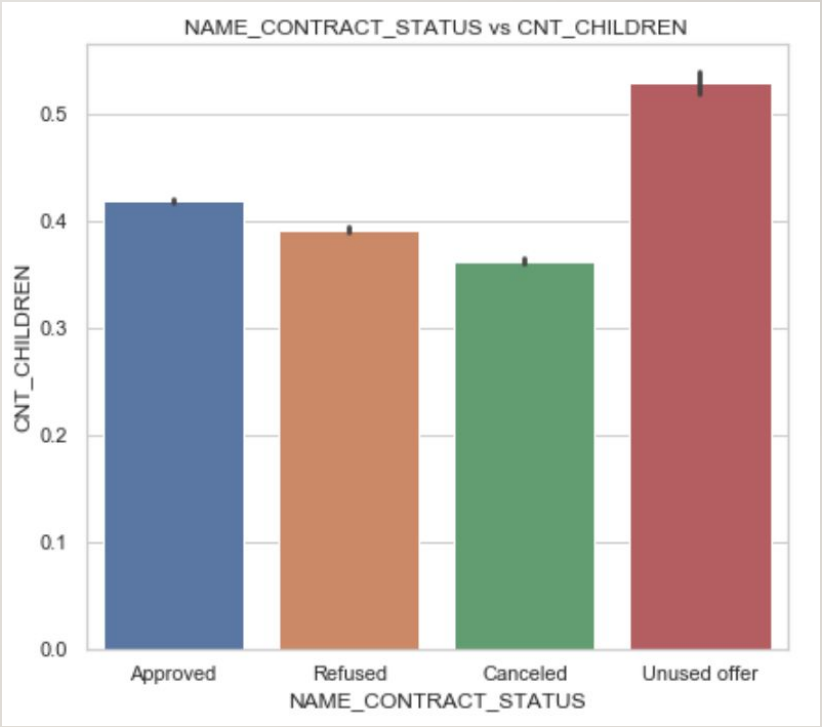
CONTINUE …..

- Count for cancelled application is highest.
- The unused offer have least application count.
- The difference between approved and refused application is ~5000.



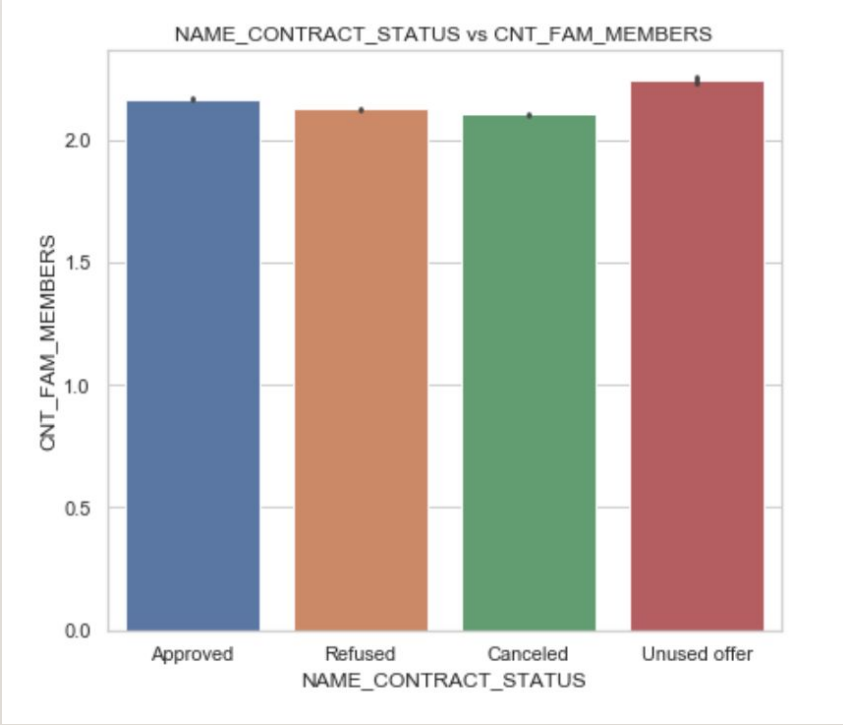
CONTINUE …..

- For maximum count child, application are under unused odder.
- For approved application, the maximum count child is 4+;
- There is slight difference for refused and cancelled application.



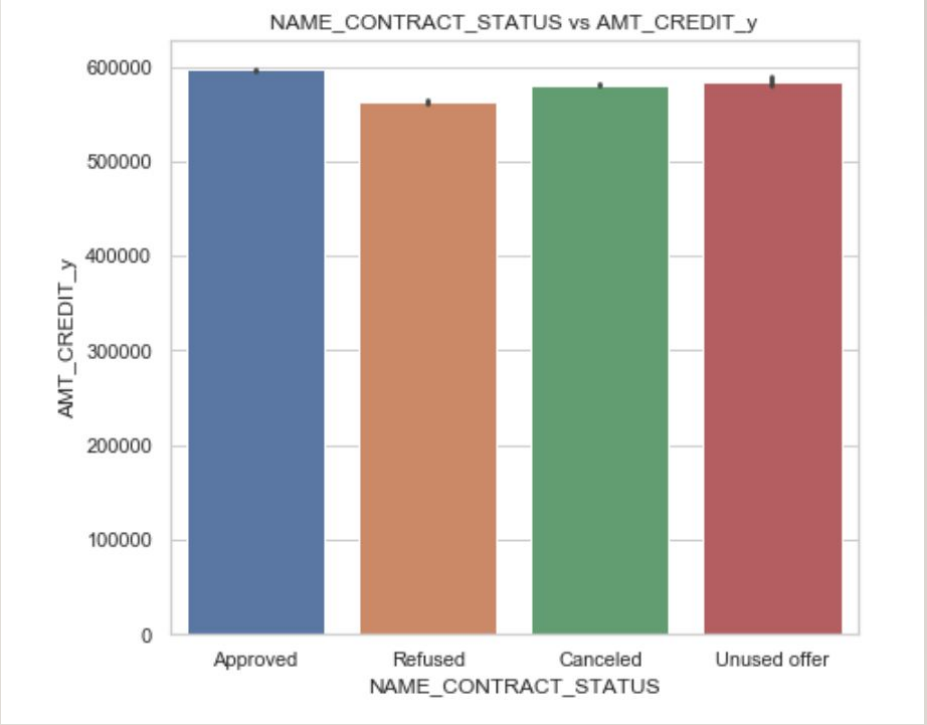
CONTINUE …..

- There is not much variance between the application status w.r.t. count of family member.
- The highest data lies under unused offer.



CONTINUE …..

- The maximum amount credit limit is 6Lc which is for approved applications.
- The count for refused, canceled and unused offer is lie between 5.5Lc to less than 6Lc.
- The canceled and unused offer count is almost equal with a slight difference to refused application.



CONCLUSION

- Application data has a total 121 columns, out of which 41 have more than 50% missing values.
- Application data has data imbalance, 90% data is for target variable 0 and rest is for target 1.
- There were outliers in application data and the date and age column had negative values which were corrected.
- There are only two types(90% cash loans & 10% revolving loans) of loan in application data while previous data had consumer loan too.
- The application date has 65% females and 35% male. Out of total female and male applicants 7% female and 10% male applicants have defaulted.
- In application data 51% are working professionals, 23% are commercial associates and 18% are pensioners. Out of total 9% working professionals, 7% working associates and 5 % pensioners have defaulted the loan.
- In application data 88% live in houses/apartments, 4.8% with parents and 3.6% live in municipal apartments. Out of total approx. 8% house/apartment applicants, 12% with parents and 8.5 % in municipal apartments have defaulted the loan.

CONTINUE …..

- 71% data belong to education type Secondary / secondary special, 24% Higher education and 3% Incomplete higher. Out of total 9% have education type Secondary / secondary special, 5% have Higher education and 8% have Incomplete higher education who default the loan.
- Approx. 63% applicants are married, 15% are single and 9.6% have civil marriage. Out of total 7.5% married, 9.8% single and 10% civil marriage applicants do the loan default.
- Approx. 72% are in average, 20% low and 6% are in above average income group. 8% average, 8% low and 6% above average income group do the loan default.
- In application data 51% are laborers, 15% are sales staff and 13% are core staff. Out of total 10% laborers, 9.6% sales staff and 11 % drivers have defaulted the loan.
- In previous data, loan application was Approved: 62.1 % times, Cancelled: 18.9 % times, Refused: 17.4 % times and Unused offer: 1.58 % times
- 77% previous applications were repeated while only 18% were new applications.
- In the previous application, 46% applications were filed through credit & cash offices while 29.6% were filed country-wide.


Thank
You