Credit EDA Case Study

Input Files and Shape

Input Data Sets

Current Application Data Set
Shape =(307511,122)

Previous Application Data Set
Shape = (1246320,26)

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 , had considered both files and performed necessary actions.

Application Data Analysis

Data Analysis - Missing Values

Application Data File – Missing Values > 45%

```
# Checking % of missing values and correspinding columns where percentage is greater thatn 45

percent_missing = round(application_df.isnull().sum()/len(application_df)*100,2)
missing_value_df = pd.DataFrame({'column_name': application_df.columns, 'percent_missing': percent_missing})
missing_value_df[missing_value_df.percent_missing > 45]
```

	column_name	percent_missing
OWN_CAR_AGE	OWN_CAR_AGE	65.99
EXT_SOURCE_1	EXT_SOURCE_1	56.38
APARTMENTS_AVG	APARTMENTS_AVG	50.75
BASEMENTAREA_AVG	BASEMENTAREA_AVG	58.52
YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_AVG	48.78
YEARS_BUILD_AVG	YEARS_BUILD_AVG	66.50
COMMONAREA_AVG	COMMONAREA_AVG	69.87
ELEVATORS_AVG	ELEVATORS_AVG	53.30
ENTRANCES_AVG	ENTRANCES_AVG	50.35
FLOORSMAX_AVG	FLOORSMAX_AVG	49.76
FLOORSMIN_AVG	FLOORSMIN_AVG	67.85
LANDAREA_AVG	LANDAREA_AVG	59.38
LIVINGAPARTMENTS_AVG	LIVINGAPARTMENTS_AVG	68.35
LIVINGAREA_AVG	LIVINGAREA_AVG	50.19

```
missing_value_df[missing_value_df.percent_missing > 45].count()
```

```
column_name        49
percent_missing    49
dtype: int64
```

There are 49 features having NULL values in application data set.

Handling Missing Values – Few Examples

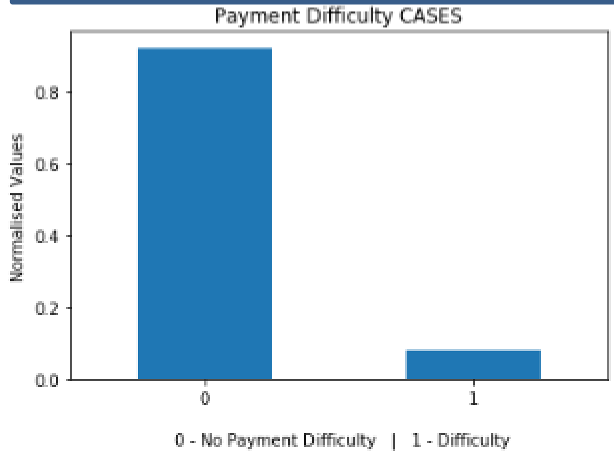
Features from 44 to 90 have data pertaining to details of living apartment / house of the applicant. These column has > 45% missing values . Hence these features are dropped from the application data set

EXT_SOURCE_1,EXT_SOURCE_2,EXT_SOURCE_3 are normalized scores received from external data source. out of these three sources, we had received 95 % of data from source 2. and other sources are having more null values. hence keeping only EXT_SOURCE_2 DATA.

All observations with AMT_GOODS_PRICE NaN is for NAME_CONTRACT_TYPE - "Revolving Lons" . Revolving loans are GENERALLY not for purchasing any partiuclar item. Hence these values NaN converted 0

Univariate Analysis

TARGET Feature

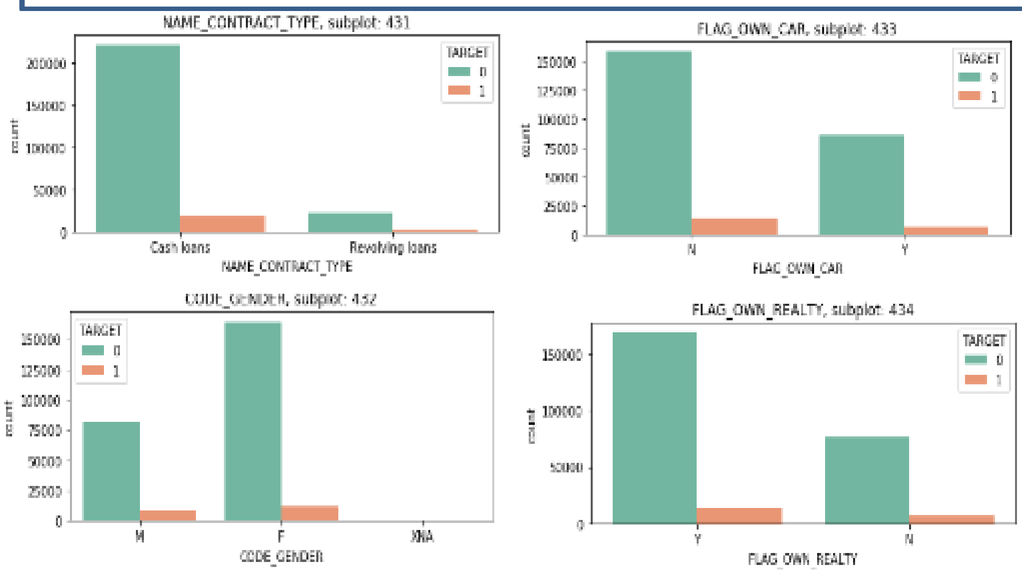



```
#checking exact Target 0 to Target 1 ratio
application_df[application_df.TARGET==0].shape[0]/application_df[application_df.TARGET==1].shape[0]
```

```
11.954366142307505
```

Inference : 1 in every ~12 applicant has payment difficulty. DATA IMBALANCE DETECTED

Object Data Type Feature Analysis – Few Examples



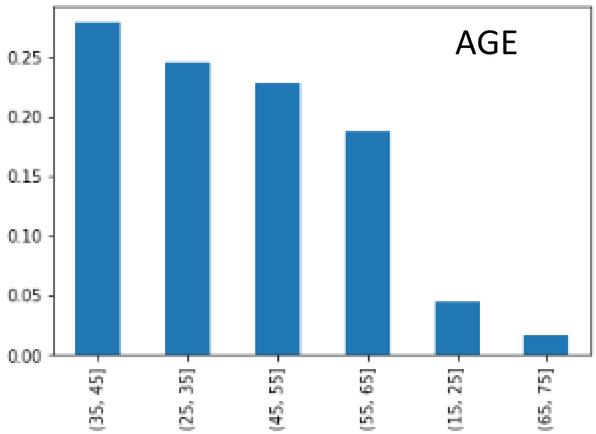
Few notable points – examples

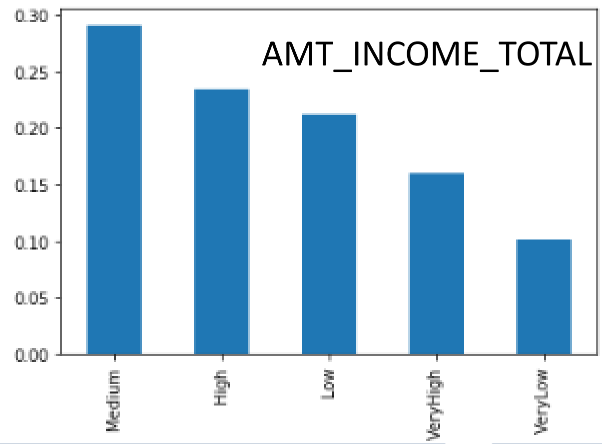
Performed countplot analysis on object features and observed below points

- Cash loans offered are more than revolving loans, at 90%
- 65% Females have taken loans in comparison to 34% male. This is very interesting and needs to be studied further
- 65% applicant don't own cars
- 69% applicants own living quarters

Univariate Analysis

Binning AGE & AMT_INCOME_TOTAL



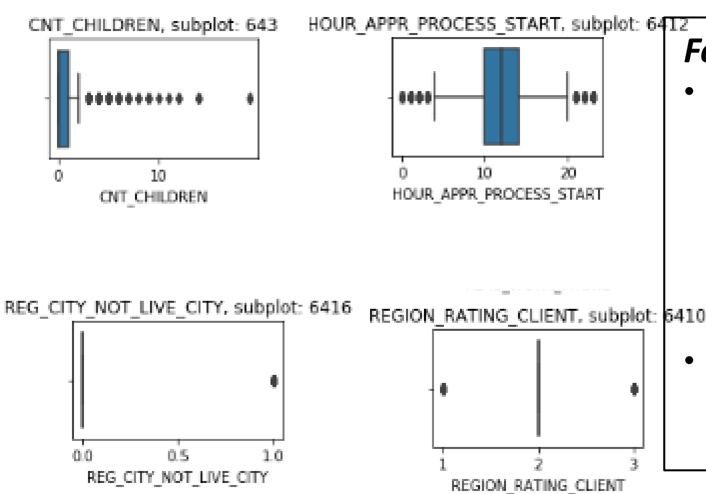


Few notable points

After Binning AGE and AMT_INCOME_TOTAL feature , applied BAR plot and observed below points

- 35-45 Age group is the largest Group of Age applying for loans. This may be attributed to consumerism aspect at that age
- Medium Income group is the largest Group applying for loans.

BOX Plot Analysis on Integer data type Features



Few notable points

- Many columns with int data type are Flag columns. For purpose of calculations we will keep them as int. eg, REG_CITY_NOT_LIVE_CITY.
- CNT_CHILDREN needs to further analysed as it has outliers

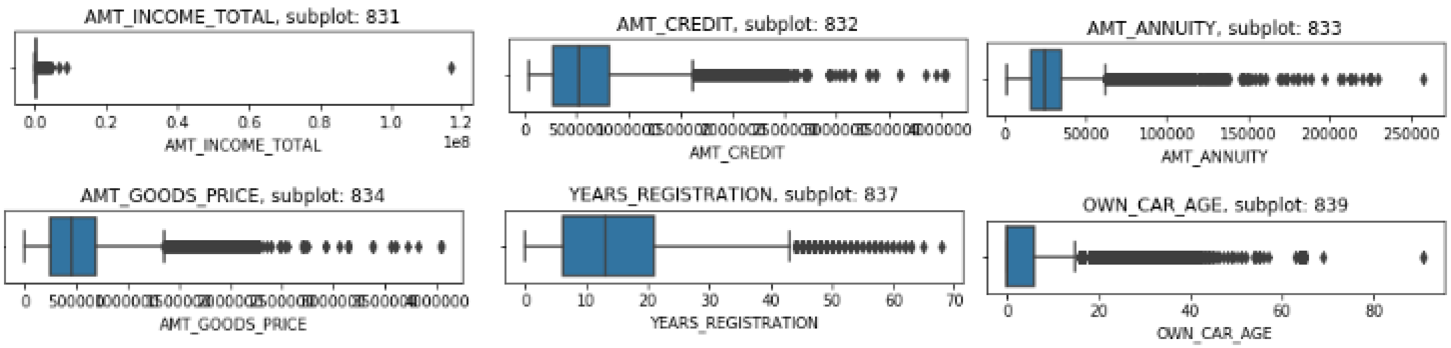
CNT_CHILDREN

```
application_df['CNT_CHILDREN'].value_counts()

0     185323
1      53362
2      23583
3       3258
4        355
5         76
6         16
7          6
14         3
19         2
12         2
9          2
8          2
11         1
10         1
```

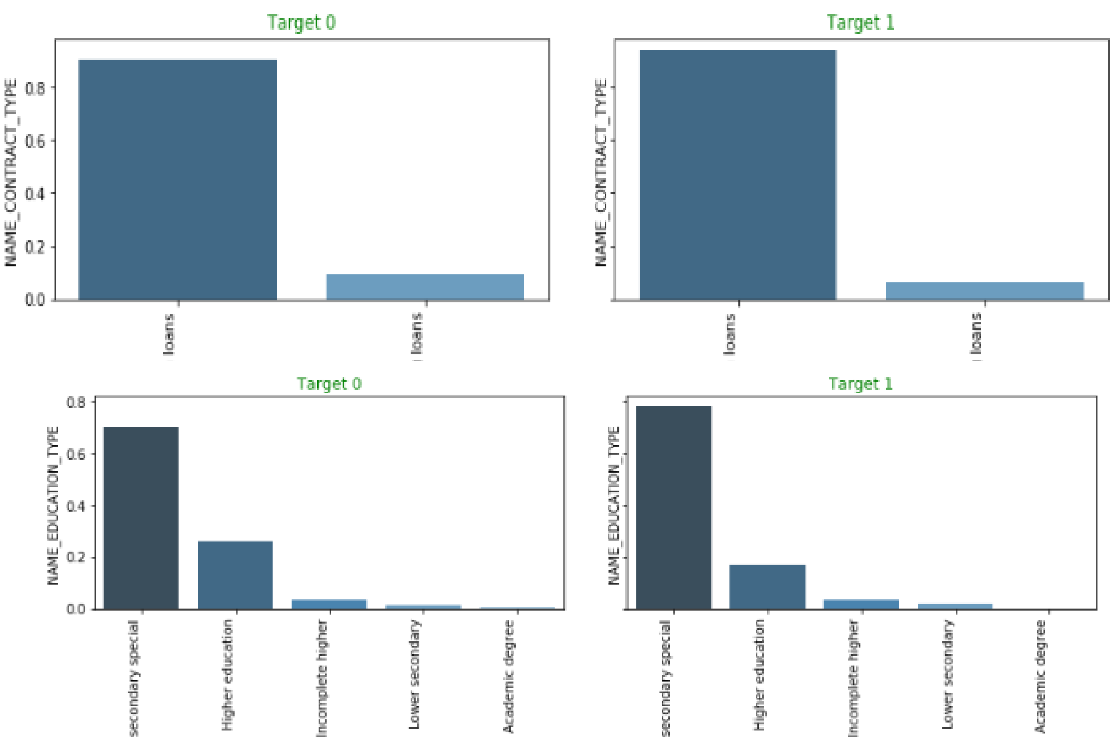
- 13 records have CNT_CHILDREN >7. These could be a possibility for outlier. Handling outlier is not a mandatory step, hence not performed any treatment for outliers.

BOX Plot Analysis on Float data type Features



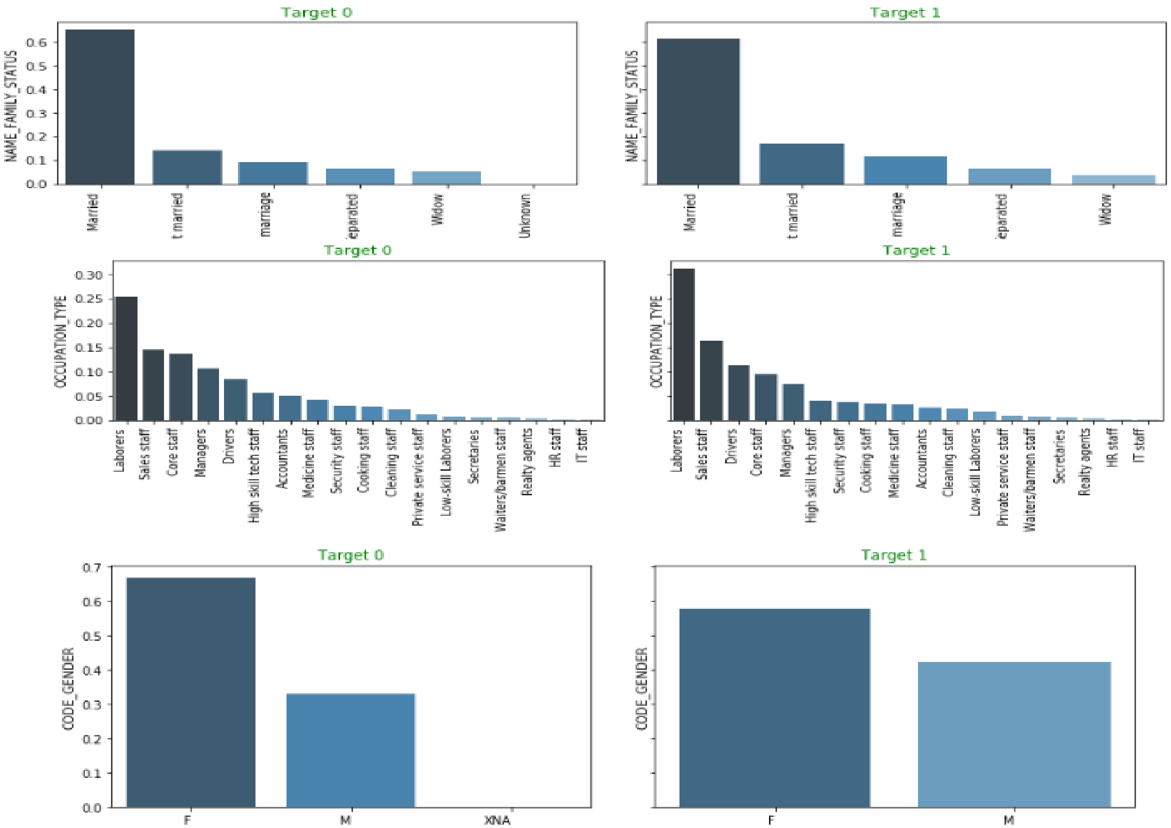
- Here are the few outliers listed post box plot analysis on Float data type features. We can substitute features with Median values. However not performing the outlier treatment as not a mandatory step.

Data Analysis – For Categorical Variables.



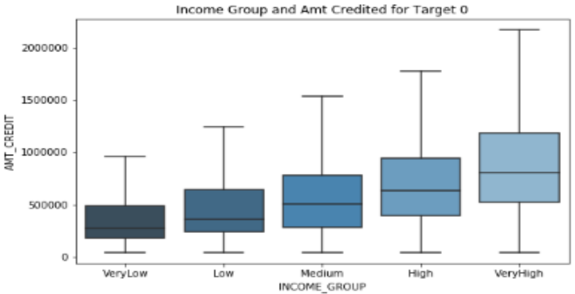
- Divided the dataset into two subsets based on Target variable. i.e. Target=0 and Target=1
- Perform Univariate analysis for categorical variables for both 0 and 1
- NAME_CONTRACT TYPE- Cash Loans are large part of the company's portfolio. For Target 0 - 85% and almost 95% for Target-1.
- NAME_EDUCATION_TYPE - In both Target 0 and 1, applicants with Secondary Education has applied for loans more than others.90% of defaulting payments are from applicants with secondary income. Needs further analysis

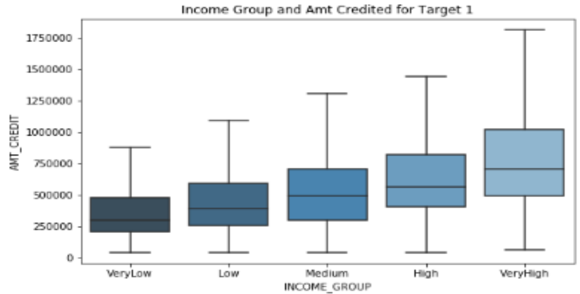
Data Analysis - – For Categorical Variables.

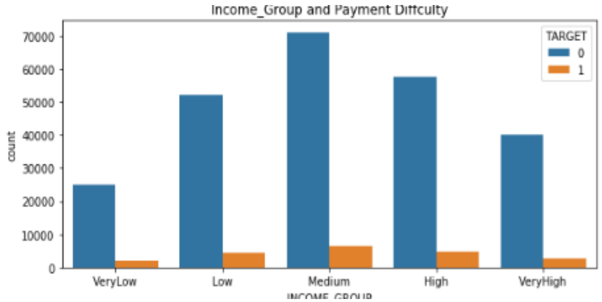


- NAME_FAMILY_STATUS - Married applicants - almost 60% have defaulted on payments
- OCCUPATION_TYPE - Labourers, sales staff, core staff, drivers constitute of 50% of defaulters. Labourers is the highest percentage of applicants too.
- CODE_GENDER' - Ratio of F to M in Target 0 is 2.3 and F to M in Target 0 - 1.3. indicating that MEN are defaulting more than Women
- **Had performed the similar analysis on all the Categorical columns.**

Bivariate Analysis on Categorical and Continues Variables.







TARGET	0	1
INCOME_GROUP		
VeryLow	351750.53	370108.92
Low	459826.05	454261.40
Medium	576062.56	549404.83
High	703181.56	647939.46
VeryHigh	890294.24	790405.74

- We can infer that though the maximum no of loans is given to Medium income group. Default value per loan is highest in High income group as the AMT_CREDIT is higher too. The loan book of the financial institution can get affected due to higher amount not being paid back.
- The company must devise a different set of rules and policies while approving higher income group loans..

TOP Correlations

Target - 0

	Column1	Column2	Correlation
474	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998528
166	AMT_GOODS_PRICE	AMT_CREDIT	0.986852
326	CNT_FAM_MEMBERS	CNT_CHILDREN	0.880158
502	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.858447
167	AMT_GOODS_PRICE	AMT_ANNUITY	0.777909
139	AMT_ANNUITY	AMT_CREDIT	0.773174
710	YRS_AGE	YEARS_EMPLOYED	0.626117
138	AMT_ANNUITY	AMT_INCOME_TOTAL	0.446181
165	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.371678
111	AMT_CREDIT	AMT_INCOME_TOTAL	0.365239

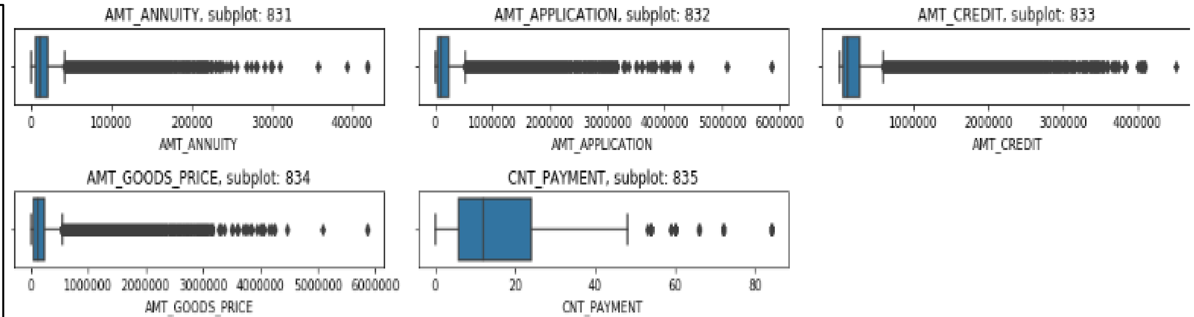
Target - 1

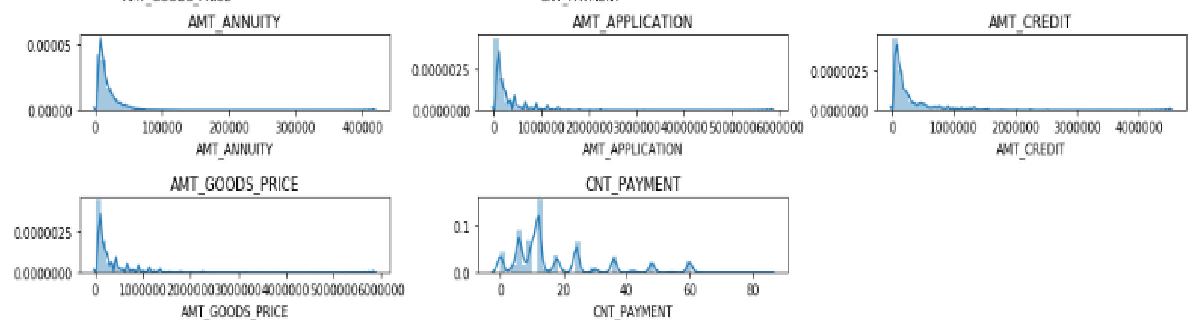
	Column1	Column2	Correlation
474	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998528
166	AMT_GOODS_PRICE	AMT_CREDIT	0.986852
326	CNT_FAM_MEMBERS	CNT_CHILDREN	0.880158
502	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.858447
167	AMT_GOODS_PRICE	AMT_ANNUITY	0.777909
139	AMT_ANNUITY	AMT_CREDIT	0.773174
710	YRS_AGE	YEARS_EMPLOYED	0.626117
138	AMT_ANNUITY	AMT_INCOME_TOTAL	0.446181
165	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.371678
111	AMT_CREDIT	AMT_INCOME_TOTAL	0.365239

Previous Application – Data Analysis Observations

- Had performed similar exercise on Previous application data type and drawn below observations :

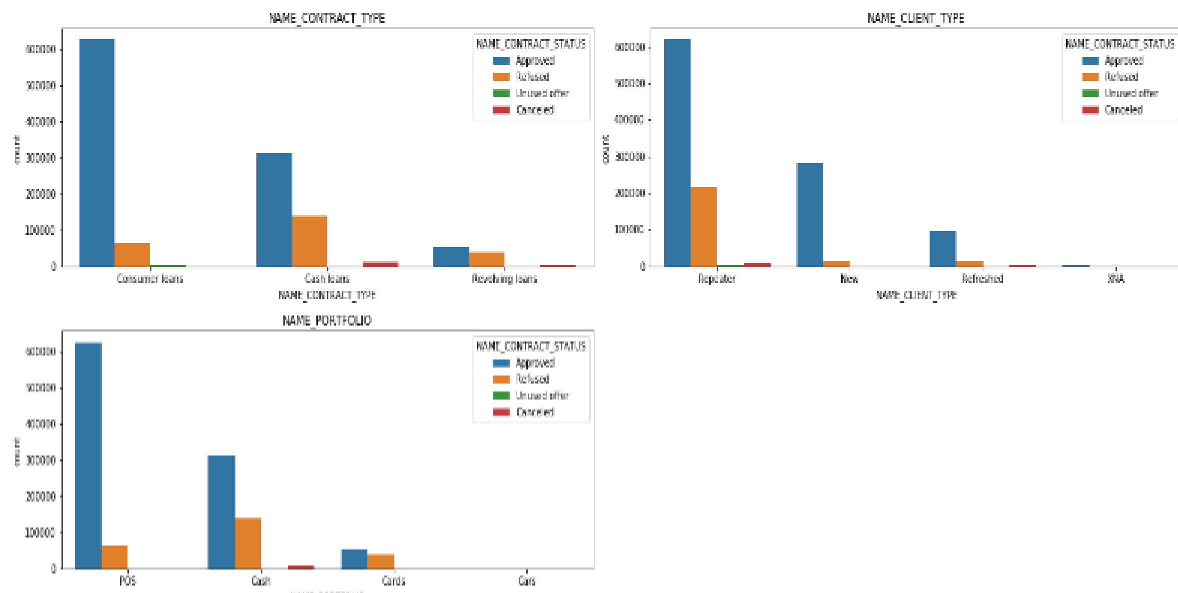
1. Continuous Variables seem to have high percentage of outliers. Checking distribution. Box plot and distribution both signify the same





2. Performed Bivariate analysis and below observations are identified

- 1. In approved category, consumer loan has largest no of applicants.
- . There seem to be no cancelled loans in cash loan category than consumer loan.
- . More cash loans have been refused than consumer loans.
- The bank has more repeaters in all approved, refused, unused, cancelled categories
- POS transactions seem to be consumer loans and similar to point 2 - more cash laons have been refused than POS.



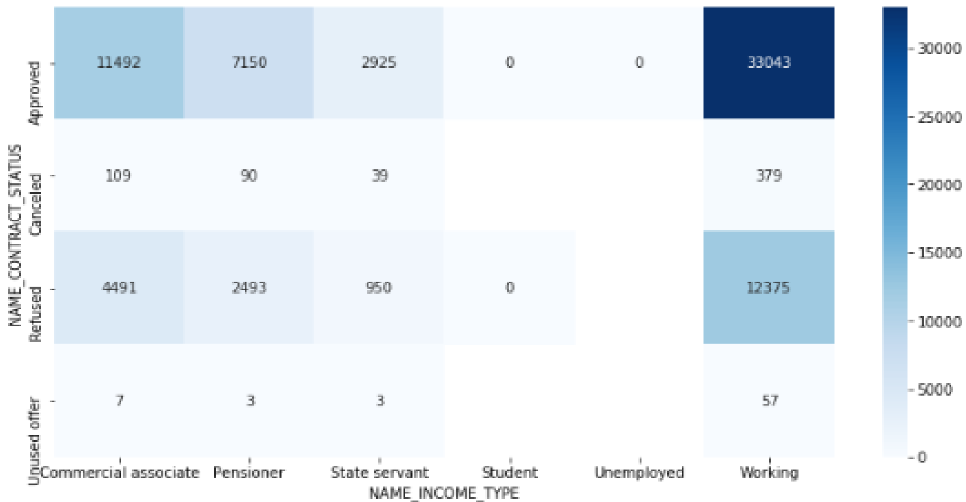
TOP Correlations – Previous application type

	Column1	Column2	Correlation
16	AMT_GOODS_PRICE	AMT_APPLICATION	0.999883
17	AMT_GOODS_PRICE	AMT_CREDIT	0.993028
11	AMT_CREDIT	AMT_APPLICATION	0.992965
15	AMT_GOODS_PRICE	AMT_ANNUITY	0.820895
5	AMT_APPLICATION	AMT_ANNUITY	0.820831
10	AMT_CREDIT	AMT_ANNUITY	0.814884
22	CNT_PAYMENT	AMT_CREDIT	0.700323
21	CNT_PAYMENT	AMT_APPLICATION	0.672276
23	CNT_PAYMENT	AMT_GOODS_PRICE	0.672129
20	CNT_PAYMENT	AMT_ANNUITY	0.401020

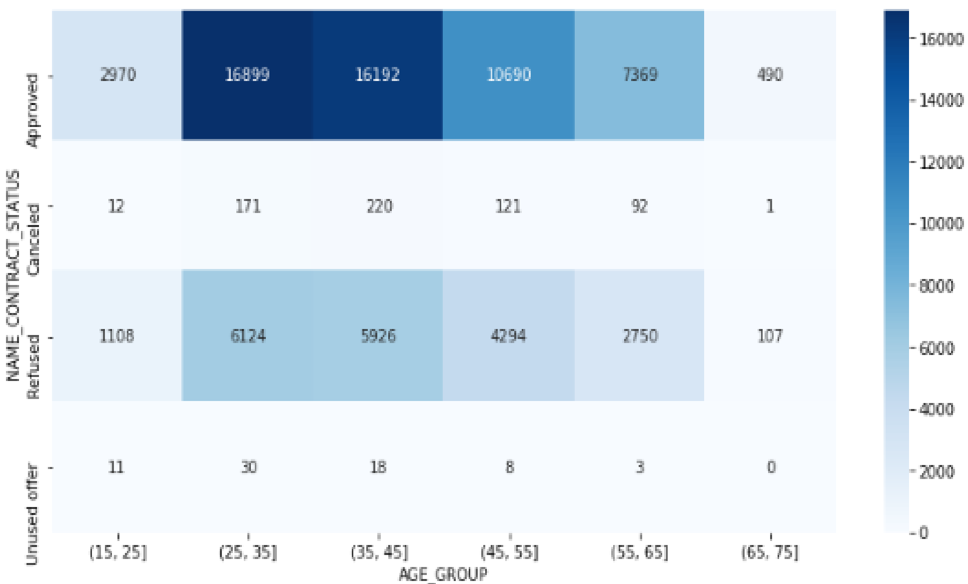
- 1. AMT_GOODS_PRICE, AMT_ANNUITY, AMT_APPLICATION - as expected have high correlation. Higher the value of good purchased more there will be need of loan and surely all these will correlate
- 2. AMT_Credit to AMT_GOOD_PRICE also the correlation is high

Merge Data Analysis

- Merge the datasets Application and previous on SK_Current_ID (inner join)



1. Since Target 1 is default, higher on the above matrix shows correlation to default.
2. Working applicant with Approved status have defaulted in highest numbers
3. Previous applications with Refused, Cancelled, Unused loans also have default which is a matter of concern. This indicates that the financial company had Refused/cancelled previous application, but has approved the current and is facing default on these loans.
4. 12,375 applicants of working class were REFUSED earlier and now have defaulted.



1. Approved loans of age group 25-35 and 35-45 have higher defaults
2. Refused, cancelled, loans in previous application have defaulted in current.

CASE STUDY SUMMARY

All the below variables were established in analysis of Application data frame as leading to default. Checked these against the Approved loans which have defaults, and it proves to be correct.

- Medium income
- 25-35 years old, followed by 35-45 years age group
- Male
- Unemployed
- Labourers, Salesman, Drivers
- Business type 3
- Own House - No

Other IMPORTANT Factors to be considered

- Days last phone number changed - Lower figure points at concern
- No of Bureau Hits in last week. Month etc – zero hits is good
- Amount income not correspondingly equivalent to Good Bought – Income low and good value high is a concern
- Previous applications with Refused, Cancelled, Unused loans

- Unused applications have lower loan amount.
- Female applicants should be given extra weightage as defaults are lesser
- 60% of defaulters are Working applicants. This does not mean working applicants must be refused. Proper scrutiny of other parameters needed
- Previous applications with Refused, Cancelled, Unused loans also have cases where payments are coming on time in current application. This indicates that possibly wrong decisions were done in those cases