CREDIT EDA CASE STUDY

Prepared by:
Priyadarshini Resapu
Silpa Devarapalli

Problem Statement:

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. we have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, and then approving the loan may lead to a financial loss for the company.

The data contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample,

All other cases: All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

Approved: The Company has approved loan Application

Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.

Refused: The company had rejected the loan (because the client does not meet their requirements etc.).

Unused offer: Loan has been cancelled by the client but on different stages of the process.

Introduction

This EDA aims at identifying key driving factors (consumer attributes and loan attributes) which contribute to defaulting a loan by a consumer. There are 2 data sets and 1 column-description file provided for the case study. They are namely:

application_data.csv

previous_application.csv

columns_description.csv

We've performed the below steps for this case study:

1. Data Understanding and preparation
2. Data Cleaning and Manipulation
3. Data Analysis
4. Presentations and Recommendations

1. Data Understanding and preparation

- First we have imported the required libraries for working with the datasets and plotting the visualizations.
- Performed data sourcing by reading the required datasets application_data (details of consumer applying for a new loan) and previous_application(loan details and customer details of existing or closed loans)
- Carried out some inspections on the application data to check the following:

  - First 5 records using head function

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CRE
0	100002	1	Cash loans	M	N	Y	0	202500.0	40659
1	100003	0	Cash loans	F	N	N	0	270000.0	129350
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	13500
3	100006	0	Cash loans	F	N	Y	0	135000.0	31268
4	100007	0	Cash loans	M	N	Y	0	121500.0	51300

  - Number of rows and columns

```
(Rows,Columns) : (307511, 122)
```

  - Data types and missing values column wise

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 122 columns):
 #   Column                       Non-Null Count   Dtype
---  ------                       --------------   -----
 0   SK_ID_CURR                   307511 non-null  int64
 1   TARGET                       307511 non-null  int64
 2   NAME_CONTRACT_TYPE           307511 non-null  object
 3   CODE_GENDER                  307511 non-null  object
 4   FLAG_OWN_CAR                 307511 non-null  object
 5   FLAG_OWN_REALTY              307511 non-null  object
 6   CNT_CHILDREN                 307511 non-null  int64
 7   AMT_INCOME_TOTAL             307511 non-null  float64
 8   AMT_CREDIT                   307511 non-null  float64
 9   AMT_ANNUITY                  307499 non-null  float64
 10  AMT_GOODS_PRICE              307233 non-null  float64
 11  NAME_TYPE_SUITE              306219 non-null  object
 12  NAME_INCOME_TYPE             307511 non-null  object
 13  NAME_EDUCATION_TYPE          307511 non-null  object
 14  NAME_FAMILY_STATUS           307511 non-null  object
 15  NAME_HOUSING_TYPE            307511 non-null  object
```

- Checked Null value counts and percentage of null values for each column in the data frame

	Null Count	Null Percentage
COMMONAREA_MEDI	214865	69.87
COMMONAREA_AVG	214865	69.87
COMMONAREA_MODE	214865	69.87
NONLIVINGAPARTMENTS_MODE	213514	69.43
NONLIVINGAPARTMENTS_AVG	213514	69.43
NONLIVINGAPARTMENTS_MEDI	213514	69.43
FONDKAPREMONT_MODE	210295	68.39
LIVINGAPARTMENTS_MODE	210199	68.35
LIVINGAPARTMENTS_AVG	210199	68.35
LIVINGAPARTMENTS_MEDI	210199	68.35
FLOORSMIN_AVG	208642	67.85
FLOORSMIN_MODE	208642	67.85

- Dropped the attributes where missing value % >= 45 as they seem to have limited significance in the present analysis. Again checked the missing value percentage for all the columns.

	Null Count	Null Percentage
OCCUPATION_TYPE	96391	31.35
EXT_SOURCE_3	60965	19.83
AMT_REQ_CREDIT_BUREAU_YEAR	41519	13.50
AMT_REQ_CREDIT_BUREAU_QRT	41519	13.50
AMT_REQ_CREDIT_BUREAU_MON	41519	13.50
AMT_REQ_CREDIT_BUREAU_WEEK	41519	13.50
AMT_REQ_CREDIT_BUREAU_DAY	41519	13.50
AMT_REQ_CREDIT_BUREAU_HOUR	41519	13.50
NAME_TYPE_SUITE	1292	0.42
OBS_30_CNT_SOCIAL_CIRCLE	1021	0.33
DEF_30_CNT_SOCIAL_CIRCLE	1021	0.33
OBS_60_CNT_SOCIAL_CIRCLE	1021	0.33

- Post dropping the attributes with more than 45% of missing values, we started imputing the other missing columns with NaN Percentage<=13% or whichever necessary.
- **OCCUPATION_TYPE** - In order to prevent data loss and also perform good analysis, dealt with the NaN values by categorizing them into a new category "missing".

```
missing                 96391
Laborers                55186
Sales staff             32102
Core staff              27570
Managers                21371
Drivers                 18603
High skill tech staff   11380
Accountants              9813
Medicine staff           8537
Security staff           6721
Cooking staff            5946
Cleaning staff           4653
Private service staff    2652
Low-skill Laborers       2093
Waiters/barmen staff     1348
Secretaries              1305
Realty agents             751
HR staff                  563
IT staff                  526
Name: OCCUPATION_TYPE, dtype: int64
```

- **EXT_SOURCE_3** – Imputed the numerical column with median as mean is far

```
# Checking after replacement
curr_appl2.EXT_SOURCE_3.isna().sum()
```

```
0
```

- **AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT and AMT_REQ_CREDIT_BUREAU_YEAR** - Imputed all the CREDIT BUREAU EQUIRY related numerical columns with 0 assuming no enquiries made. Also, it's inappropriate to have counts as floats. Hence, converted the data type of all these columns to int.

```
curr_appl2.loc[:,"AMT_REQ_CREDIT_BUREAU_HOUR":"AMT_REQ_CREDIT_BUREAU_YEAR"].isna().sum()
```

```
AMT_REQ_CREDIT_BUREAU_HOUR    0
AMT_REQ_CREDIT_BUREAU_DAY     0
AMT_REQ_CREDIT_BUREAU_WEEK    0
AMT_REQ_CREDIT_BUREAU_MON     0
AMT_REQ_CREDIT_BUREAU_QRT     0
AMT_REQ_CREDIT_BUREAU_YEAR    0
dtype: int64
```

- **NAME_TYPE_SUITE** - Imputed the missing values with mode (Unaccompanied) for this categorical column.

```
# Rechecking the column NAME_TYPE_SUITE for missing values after imputation
curr_appl2.NAME_TYPE_SUITE.isna().sum()
```

```
0
```

- **SOCIAL SURROUNDINGS COLUMNS (OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE)** - Imputed Client's social surroundings observation counts related null values with 0 assuming there are no observations made. Also, converted the columns type to int as the counts cannot have float values

```
# Checking for null values fater the necessary changes
curr_appl2.loc[:,"OBS_30_CNT_SOCIAL_CIRCLE":"DEF_60_CNT_SOCIAL_CIRCLE"].isna().sum()
```

```
OBS_30_CNT_SOCIAL_CIRCLE    0
DEF_30_CNT_SOCIAL_CIRCLE    0
OBS_60_CNT_SOCIAL_CIRCLE    0
DEF_60_CNT_SOCIAL_CIRCLE    0
dtype: int64
```

```
OBS_30_CNT_SOCIAL_CIRCLE dtype :int32
DEF_30_CNT_SOCIAL_CIRCLE dtype :int32
OBS_60_CNT_SOCIAL_CIRCLE dtype :int32
DEF_60_CNT_SOCIAL_CIRCLE dtype :int32
```

- **EXT_SOURCE_2** - Imputed the missing values with median for this numerical column as the mean is far.

```
# Displaying the changes
curr_appl2.EXT_SOURCE_2.isna().sum()
```

```
0
```

- **AMT_GOODS_PRICE** - Imputed the missing values with median for this numerical column as the mean value is far.

```

#Check the null values in the AMT_GOODS_PRICE after imputing with median
curr_appl2.AMT_GOODS_PRICE.isnull().sum()
```

```
0
```

- **AMT_ANNUITY** - Imputed missing values with mean for the numeric column AMT_ANNUITY as there is a huge difference between the max and 75%

```
#Check the null values in the AMT_ANUITY after imputing with mean
curr_appl2.AMT_ANNUITY.isna().sum()
```

```
0
```

- **CNT_FAM_MEMBERS** - Imputed CNT_FAM_MEMBERS with 1 as all other family details are missing or 0. Also, the counts cannot be float values. Hence, converting the data type to int.

```
#Check the null values and DTYPE in the CNT_FAM_MEMBERS after necessary changes
print("MIISING VALUES:",curr_appl2.CNT_FAM_MEMBERS.isna().sum())
print("DTYPE:",curr_appl2.CNT_FAM_MEMBERS.dtype)
```

```
MIISING VALUES: 0
DTYPE: int32
```

- **DAYS_LAST_PHONE_CHANGE** - Imputed the null values of this numerical column with 0 as the mean and median are so far . Also, changed the data type of the column to int as days cannot be floats.

```
print("DTYPE:",curr_appl2.DAYS_LAST_PHONE_CHANGE.dtype)
# Checking the negative value conversion
curr_appl2["DAYS_LAST_PHONE_CHANGE"].head()
```

```
MIISING VALUES: 0
DTYPE: int32

0    1134
1     828
2     815
3     617
4    1106
Name: DAYS_LAST_PHONE_CHANGE, dtype: int32
```

- **DAYS RELATED COLUMNS (DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH) -** These columns have values in negative. So, changed them to positive values by using pd.series.abs() and converted them into integers from floats as days cannot be float values.

	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH
0	9461	637	3648	2120
1	16765	1188	1186	291
2	19046	225	4260	2531
3	19005	3039	9833	2437
4	19932	3038	4311	3458

```
DAYS_BIRTH int32
DAYS_EMPLOYED int32
DAYS_REGISTRATION int32
DAYS_ID_PUBLISH int32
```

- **CODE_GENDER -** The null value count in the column is 0 but “XNA” values clearly imply missing values. As the percentage of these "XNA" value is quite low, imputed them with mode.

```
#Checking Gender value counts afterimputing with mode.
curr_appl2.CODE_GENDER.value_counts()
```

```
F    202452
M    105059
Name: CODE_GENDER, dtype: int64
```

- Post imputing the missing/junk values in the categorical and numerical columns, checked all the object type columns if they have any duplicates and no duplicates found.

```
NAME_CONTRACT_TYPE :
 ['Cash loans' 'Revolving loans']

CODE_GENDER :
 ['M' 'F']

FLAG_OWN_CAR :
 ['N' 'Y']

FLAG_OWN_REALTY :
 ['Y' 'N']

NAME_TYPE_SUITE :
 ['Unaccompanied' 'Family' 'Spouse, partner' 'Children' 'Other_A' 'Other_B'
 'Group of people']

NAME_INCOME_TYPE :
 ['Working' 'State servant' 'Commercial associate' 'Pensioner' 'Unemployed'
 'Student' 'Businessman' 'Maternity leave']
```

```
NAME_EDUCATION_TYPE :
 ['Secondary / secondary special' 'Higher education' 'Incomplete higher'
 'Lower secondary' 'Academic degree']

NAME_FAMILY_STATUS :
 ['Single / not married' 'Married' 'Civil marriage' 'Widow' 'Separated'
 'Unknown']

NAME_HOUSING_TYPE :
 ['House / apartment' 'Rented apartment' 'With parents'
 'Municipal apartment' 'Office apartment' 'Co-op apartment']

OCCUPATION_TYPE :
 ['Laborers' 'Core staff' 'Accountants' 'Managers' 'missing' 'Drivers'
 'Sales staff' 'Cleaning staff' 'Cooking staff' 'Private service staff'
 'Medicine staff' 'Security staff' 'High skill tech staff'
 'Waiters/barmen staff' 'Low-skill Laborers' 'Realty agents' 'Secretaries'
 'IT staff' 'HR staff']
```

```
WEEKDAY_APPR_PROCESS_START :
 ['WEDNESDAY' 'MONDAY' 'THURSDAY' 'SUNDAY' 'SATURDAY' 'FRIDAY' 'TUESDAY']

ORGANIZATION_TYPE :
 ['Business Entity Type 3' 'School' 'Government' 'Religion' 'Other' 'XNA'
 'Electricity' 'Medicine' 'Business Entity Type 2' 'Self-employed'
 'Transport: type 2' 'Construction' 'Housing' 'Kindergarten'
 'Trade: type 7' 'Industry: type 11' 'Military' 'Services'
 'Security Ministries' 'Transport: type 4' 'Industry: type 1' 'Emergency'
 'Security' 'Trade: type 2' 'University' 'Transport: type 3' 'Police'
 'Business Entity Type 1' 'Postal' 'Industry: type 4' 'Agriculture'
 'Restaurant' 'Culture' 'Hotel' 'Industry: type 7' 'Trade: type 3'
 'Industry: type 3' 'Bank' 'Industry: type 9' 'Insurance' 'Trade: type 6'
 'Industry: type 2' 'Transport: type 1' 'Industry: type 12' 'Mobile'
 'Trade: type 1' 'Industry: type 5' 'Industry: type 10' 'Legal Services'
 'Advertising' 'Trade: type 5' 'Cleaning' 'Industry: type 13'
 'Trade: type 4' 'Telecom' 'Industry: type 8' 'Realtor' 'Industry: type 6']
```

2. Data Cleaning and Manipulation

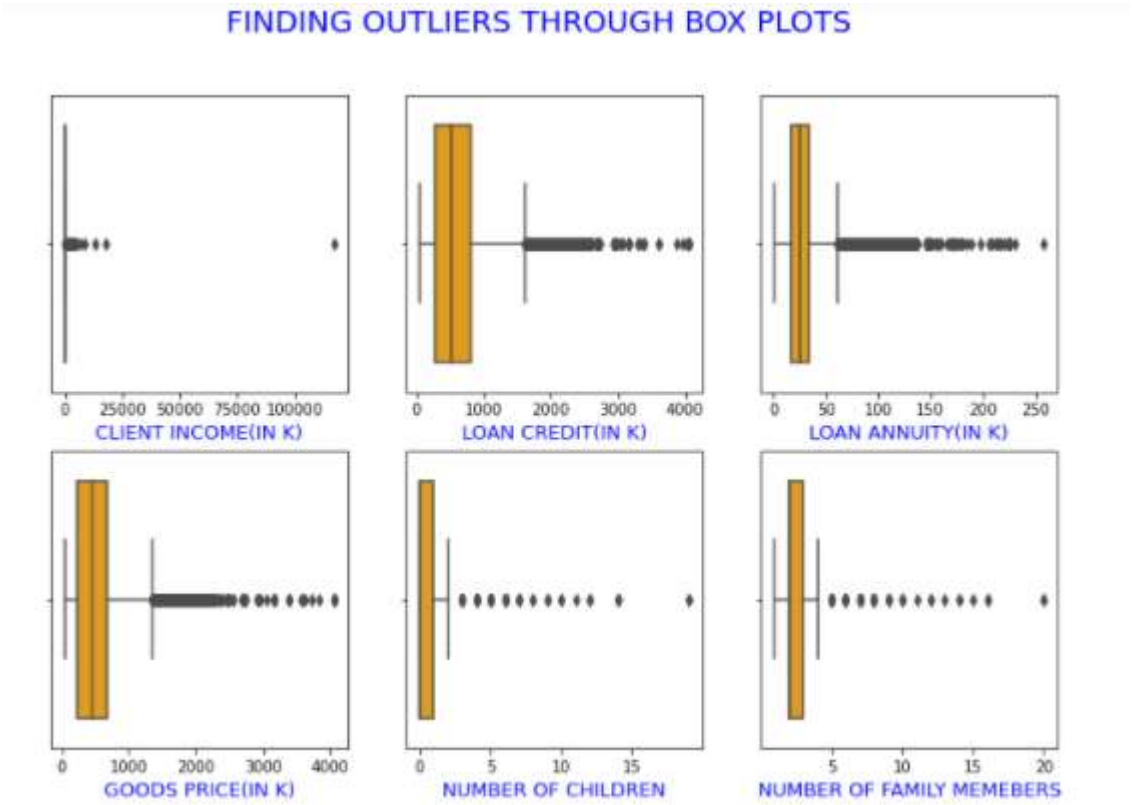
- After working on the NaN values and data type conversion, proceeded with further data cleaning and manipulation for the required columns by changing the data to standard format.
- **DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION,DAYS_ID_PUBLISH**, **DAYS_LAST_PHONE_CHANGE** - converted them into years and saved the values in the new columns with the data type int for better visualization and dropped the original columns.

	CLIENT AGE(IN YRS)	EMPLOYMENT AGE(IN YRS)	REGISTRATION AGE(IN YRS)	IDENTITY AGE(IN YRS)	CLIENT PHONE AGE(IN YRS)
count	307507.000000	307507.000000	307507.000000	307507.000000	307507.000000
mean	43.436055	185.023788	13.168712	7.713473	2.225114
std	11.954577	381.974156	9.646833	4.134528	2.193685
min	20.000000	0.000000	0.000000	0.000000	0.000000
25%	34.000000	2.000000	5.000000	4.000000	0.000000
50%	43.000000	6.000000	12.000000	8.000000	2.000000
75%	53.000000	15.000000	20.000000	11.000000	4.000000
max	69.000000	1000.000000	67.000000	19.000000	11.000000

- **AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE** - For a better visualization stored the Amount related values in thousands in new columns and then dropped the original columns.

	INCOME(IN K)	LOAN CREDIT(IN K)	LOAN ANNUITY(IN K)	GOODS PRICE(IN K)
count	307507.000000	307507.000000	307507.000000	307507.000000
mean	168.797685	599.028640	27.108708	538.317809
std	237.124628	402.492583	14.493451	369.289766
min	25.650000	45.000000	1.620000	40.500000
25%	112.500000	270.000000	16.520000	238.500000
50%	147.150000	513.530000	24.900000	450.000000
75%	202.500000	808.650000	34.600000	679.500000
max	117000.000000	4050.000000	258.030000	4050.000000

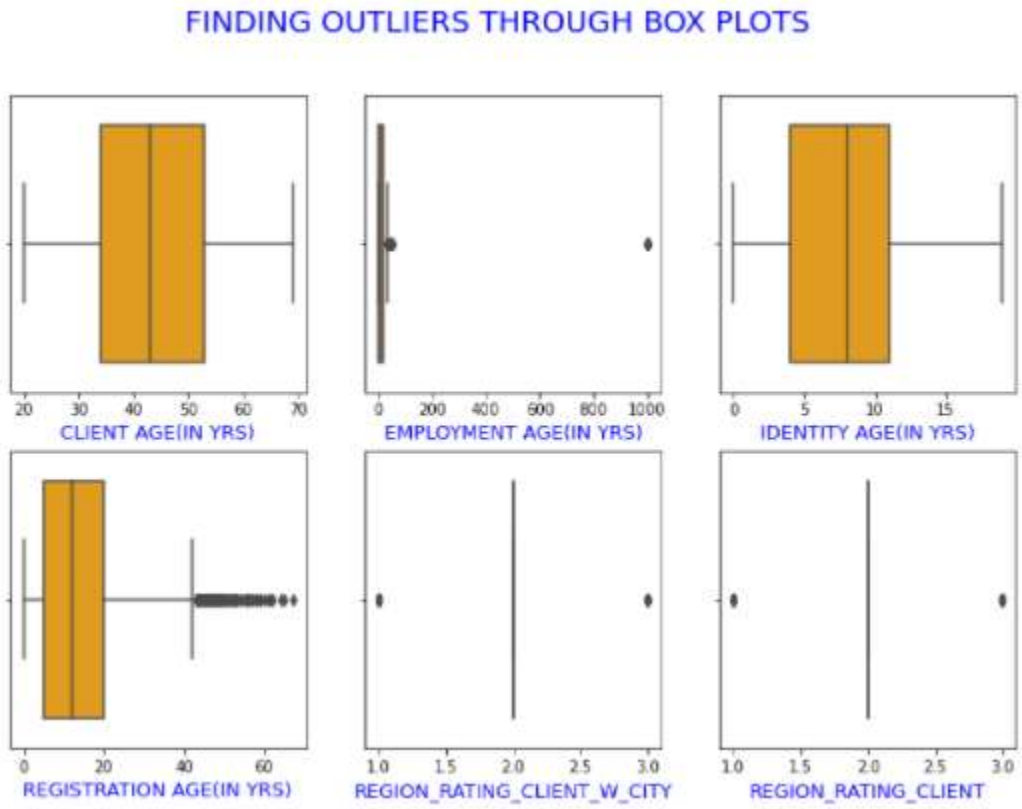
- Checked for outliers in the numeric columns **INCOME(IN K), LOAN CREDIT(IN K), LOAN ANNUITY(IN K), GOODS PRICE(IN K),CNT_CHILDREN, CNT_FAM_MEMBERS** using box plots as below:



OBSERVATIONS:

- The Box plots for all the columns clearly show presence of possible outliers in the data. Further in the data summary also the max values lie high above the 75 % indicating the presence of outliers.
- For the column **INCOME(IN K),** the max value(117000) lies very high above the 75%(202) suggesting many outliers which could hinder statistical analysis. So, to save data as well as perform good statistical analysis it is best to impute these values with median or capping them to 99% before performing any analysis.
- The **LOAN CREDIT(IN K),LOAN ANNUITY(IN K)** columns also have high max values(4050 and 258 respectively) to their 75%(808 and 34 respectively).After further insights capping the data or imputing with mean would be apt.
- The summary **CNT_CHILDREN** show possible outlier values the maximum being 19. There can be clients with more children so for further insights the marital status, number of family members, their age can be looked into to see if these have to be deleted or imputed with the median.
- The same goes with **CNT_FAM_MEMBERS**, though many families range from 2 to 5 there are cases where the clients are living with parents and siblings in a joint family mounting up the count. So, after further insights these values could be imputed with the median of the data.

- Looked through the remaining numeric **columns CLIENT AGE(IN YRS), EMPLOYMENT AGE(IN YRS), REGISTRATION AGE(IN YRS), IDENTITY AGE(IN YRS), REGION_RATING_CLIENT_W_CITY, REGION_RATING_CLIENT** to check if there are any outliers by plotting box plots.



OBSERVATIONS:

- The box plots for **CLIENT AGE(IN YRS) and IDENTIFY AGE(IN YRS)** do not show any outliers.
- The box plot for **EMPLOYMENT AGE(IN YRS)** clearly shows an outlier with age 1000 years. So, this can be excluded while doing the analysis by capping to 99%.
- **REGISTRATION AGE(IN YRS)** shows some outliers. After further analysis these Outliers can be dealt by a. Capping them with required quartile b. Imputing with mean or median whichever is apt
- The box plots for **REGION_RATING_CLIENT_W_CITY and REGION_RATIING_CLIENT** show that there are only 3 values (1, 2 and 3). These values depend upon their ratings numbered 1,2 and 3.So converting them to ordered categorical data will give better insights.

- **BINNING OF COLUMNS:** For better analysis, binned some of the continuous variables.
  - **ASSETS** - We grouped the clients into 4 categories based on their assets like House and Car namely as below:
    # "House-Car" : having both house and car
    # "House" : having house only
    # "Car" : Having car only

"No ASSETS" : No ASSETS(No house and no car)

```
curr_appl2.ASSETS.value_counts()
```

```
House        140952
House-Car     72360
No ASSETS     61972
Car           32227
Name: ASSETS, dtype: int64
```

- **INCOME(IN K)** - Binned the continuous numerical column INCOME(IN K) into 4 groups using quartiles. We've capped the records to 99% to exclude the outliers from analysis which will hinder the plots and created a new column with this data.

```
Low          100578
High          82213
Very High     68524
Medium        53182
Name: INCOME_GROUP, dtype: int64
```

- **CLIENT AGE** - Binned the continuous numerical column as per the min and max values and created a new column called **AGE GROUP**

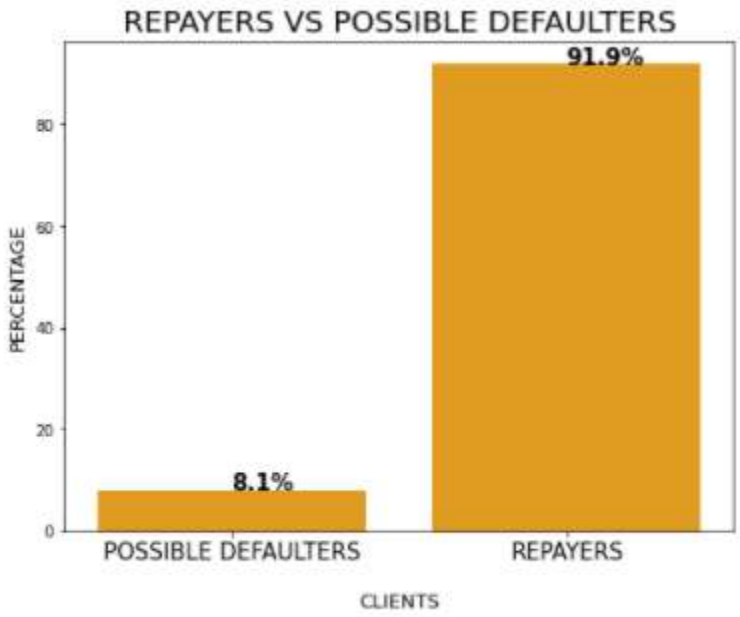
```
(30, 40]    83117
(40, 50]    74401
(50, 60]    67819
(20, 30]    52805
(60, 70]    29368
Name: AGE GROUP, dtype: int64
```

- **REGION_RATING_CLIENT_W_CITY, REGION_RATING_CLIENT** – Categorizing the values into 3 categories like tier 1, tier 2 and tier 3

```
REGION_RATING_CLIENT_W_CITY
 tier 2    229484
tier 3     43860
tier 1     34167
Name: REGION_RATING_CLIENT_W_CITY, dtype: int64

REGION_RATING_CLIENT
 tier 2    226984
tier 3     48330
tier 1     32197
Name: REGION_RATING_CLIENT, dtype: int64
```

3. Data Analysis

- The TARGET variable in the dataset has two categories represented with values 1 and 0 in which 1 Refers to clients with payment difficulties and 0 refers to all other clients. We've plotted the Target variable distribution as below:

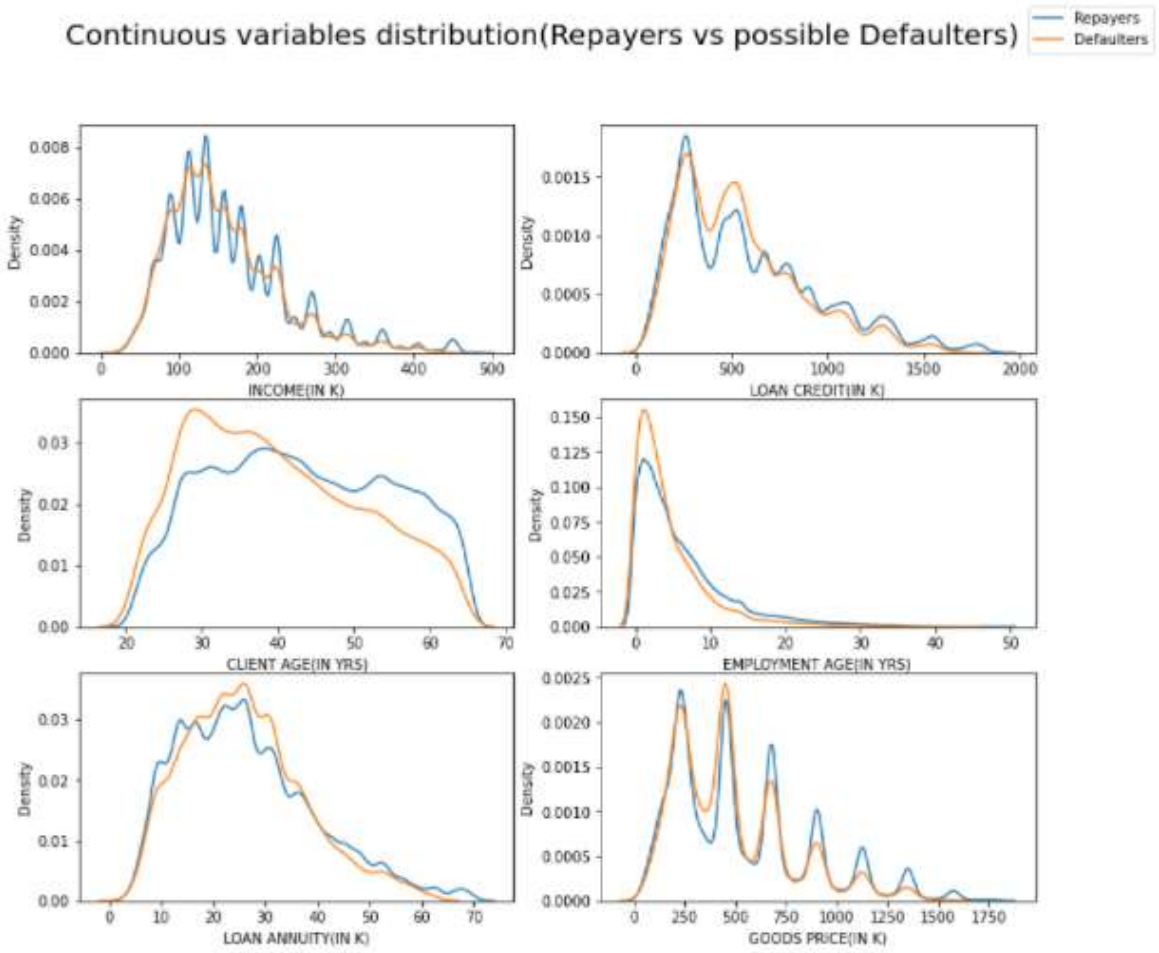


OBSERVATIONS:

- It is evident from the graph that almost 91.9% of clients are able to repay the loan on time. Clients who may be possible defaulters are very low. We have to further analyze and drill down the data in order to derive insights on what factors contribute towards a client defaulting on a loan.

- **SUBSETTING THE DATASET:** We've segregated the data set into two sub sets based on the TARGET variable. By doing so, we can observe different trends or variable patterns on both the target categories at a time independently and arrive at the key driving factors contributing to a customer being a possible defaulter.

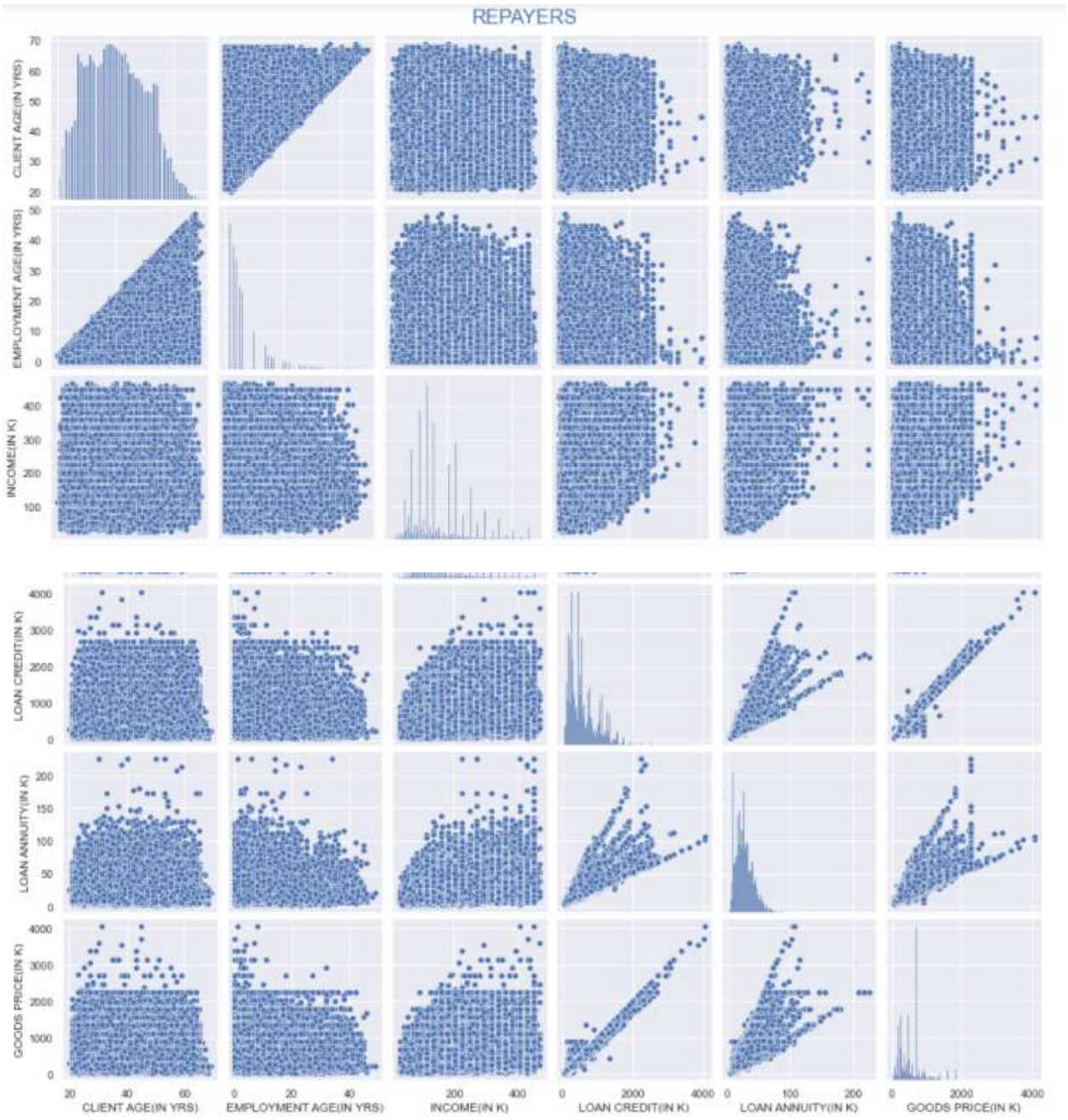
- Initially we performed **UNIVARIATE ANYLYSIS** of **REPAYERS** and **DEFAULTERS** data sets by plotting the distribution plots for the numerical columns **INCOME(IN K), LOAN CREDIT(IN K), CLIENT AGE(IN YRS), EMPLOYMENT AGE(IN YRS), LOAN ANNUITY(IN K)** and **GOODS PRICE(IN K))** as below:

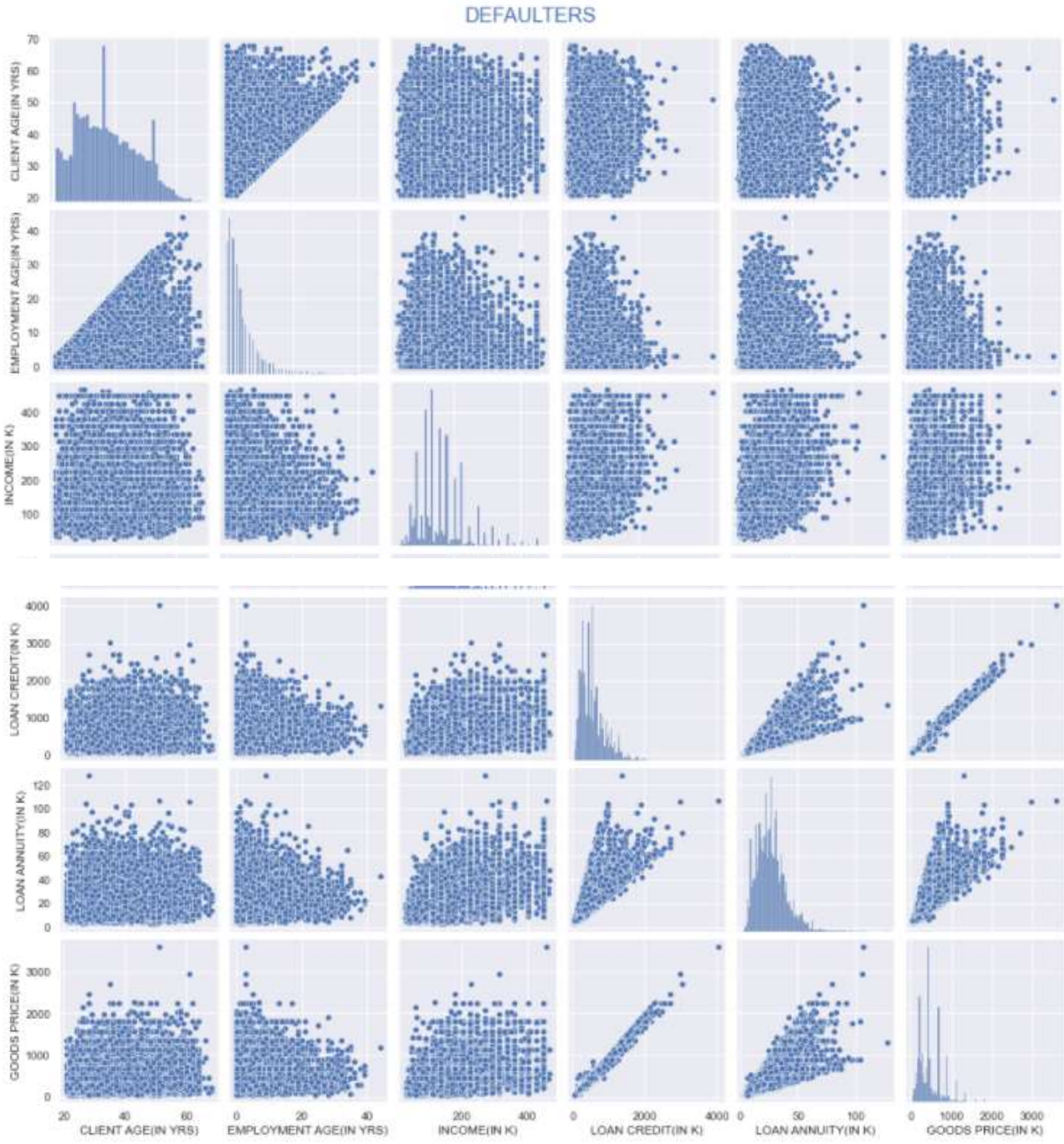


OBSERVATION:

- All the parameters seem to follow a similar distribution except for age where younger people (20-40) have higher default rate than older people (50-60).

- Then performed **BIVARIATE analysis** for the numerical columns **CLIENT AGE(IN YRS), EMPLOYMENT AGE(IN YRS), INCOME(IN K), LOAN CREDIT(IN K), LOAN ANNUITY(IN K) and GOODS PRICE(IN K)** in **REPAYERS** and **DEFAULTERS** data sets using pair plots as below:

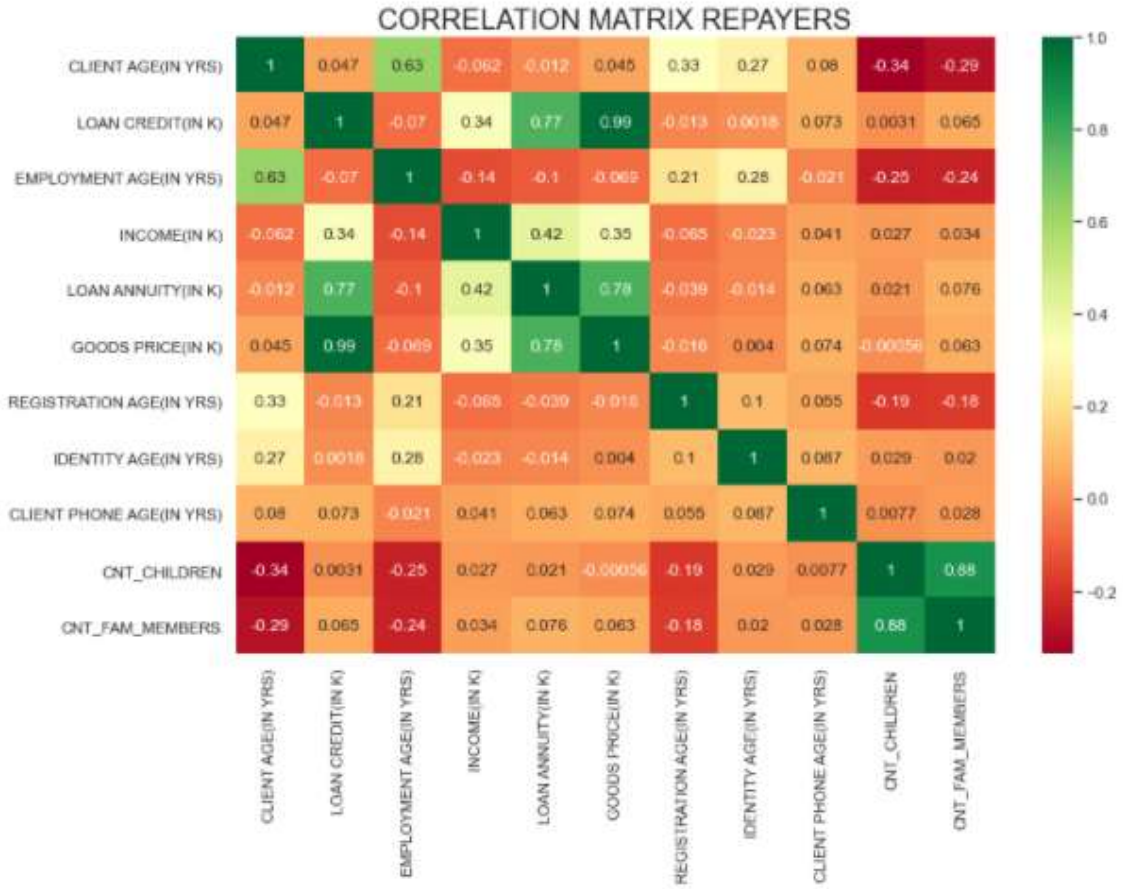

REPAYERS
CLIENT AGE(IN YRS)
EMPLOYMENT AGE(IN YRS)
INCOME(IN K)
LOAN CREDIT(IN K)
LOAN ANNUITY(IN K)
GOODS PRICE(IN K)

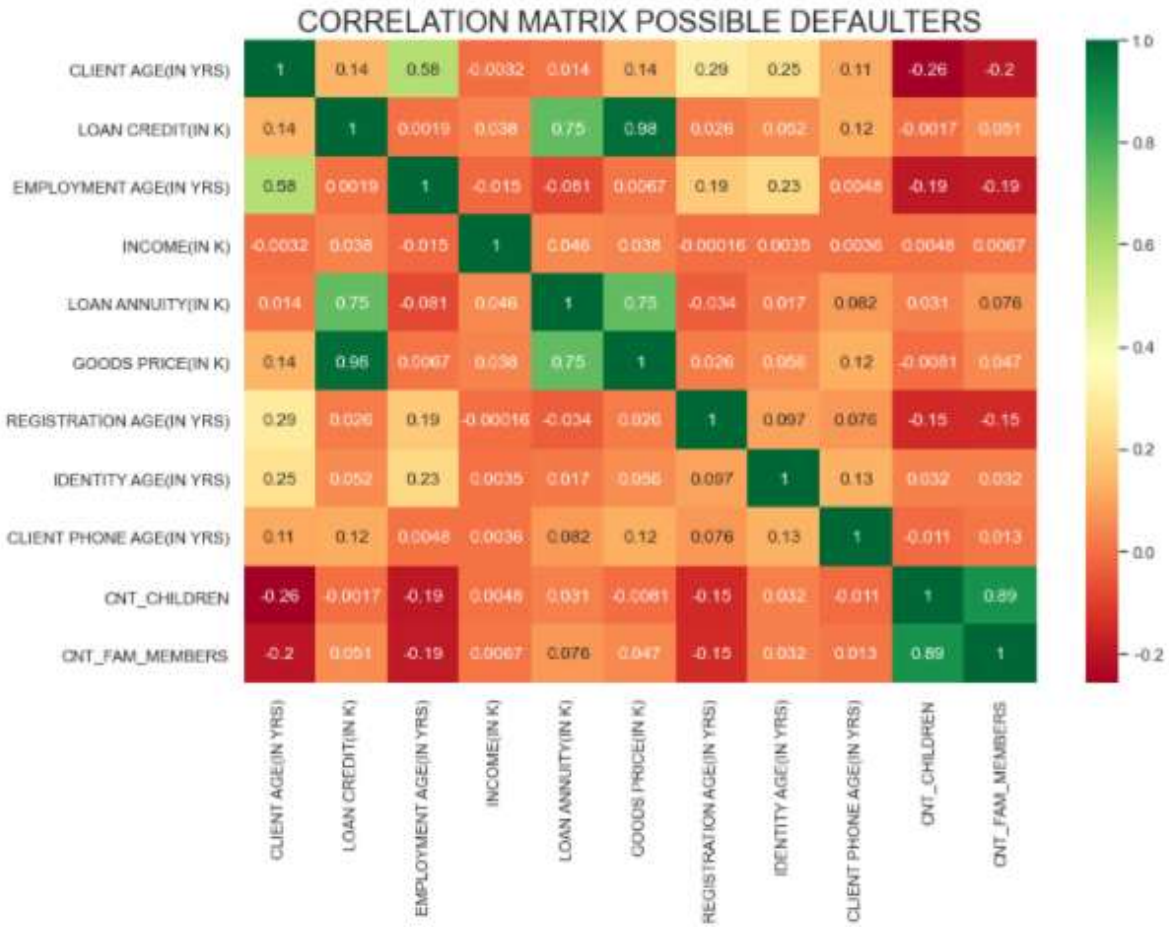


OBSERVATIONS:

- It is evident from the plots above that there is a strong linear relation between LOAN_CREDIT, GOODS_PRICE and LOAN_ANNUITY: 1. An increase in Goods price would increase the Loan Credit 2. As Loan credit increases the Annuity Amount also increases
- Due to presence of outlier values high over the 75% quartile, the visualization of INCOME, EMPLOYMENT AGE are being binned to extremes .Capping the values to 99% will give a better output.

- Performed correlation analysis for the numerical columns **CLIENT AGE(IN YRS), LOAN CREDIT(IN K), EMPLOYMENT AGE(IN YRS), INCOME(IN K), LOAN ANNUITY(IN K), GOODS PRICE(IN K), REGISTRATION AGE(IN YRS), IDENTITY AGE(IN YRS), CLIENT PHONE AGE(IN YRS), CNT_CHILDREN, CNT_FAM_MEMBERS** by plotting correlation matrix as below:





OBSERVATIONS:

- As observed from the correlation matrix there is a strong correlation between Loan credit, Goods Price and Annuity in both the groups Repayers and Possible defaulters. For both groups strong correlation exists between number of children and number of family members indicating that the family count increases if children count increases. Also, as the client age increases the employment age increases.
- In the Repayers group there exists a better correlation between the Client Income, Loan Credit, Loan Annuity and Goods price but that is not the case with Defaulters group where the correlation seems to be very weak when compared to Repayers. This seems to indicate that most defaulters are those who have been sanctioned higher loans compared to their income.

TOP TEN CORRELATIONS IN REPAYERS:

1. GOODS PRICE(IN K) and LOAN CREDIT(0.99)

2. CNT_FAMILY_MEMEBERS and CNT_CHILDREN(0.88)

3. GOODS_PRICE AND LOAN ANNUITY (0.78)

4. LOAN ANNUITY AND LOAN CREDIT(0.77)

5. EMPLOYEMENT AGE AND CLIENT AGE(0.63)

4. LOAN ANNUITY AND INCOME(0.42)

5. GOODS PRICE AND INCOME (0.35)

6. INCOME AND LOAN CREDIT (0.34)

7. CLIENT AGE AND REGISTRATION AGE(0.33)

8. IDENTITY AGE AND EMPLOYEMENT AGE(0.28)

9. CLIENT AGE AND IDENTITY AGE(0.27)

10. REGISTRATION AGE AND EMPLOYEMENT AGE(0.21)

- **TOP TEN CORRELATIONS IN DEFAULTERS:**

1. GOODS PRICE(IN K) and LOAN CREDIT(0.98)

2. CNT_FAMILY_MEMEBERS and CNT_CHILDREN(0.89)

3. GOODS_PRICE AND LOAN ANNUITY (0.75)

4. LOAN ANNUITY AND LOAN CREDIT(0.75)

5. EMPLOYEMENT AGE AND CLIENT AGE(0.58)

6. CLIENT AGE AND REGISTRATION AGE(0.29)

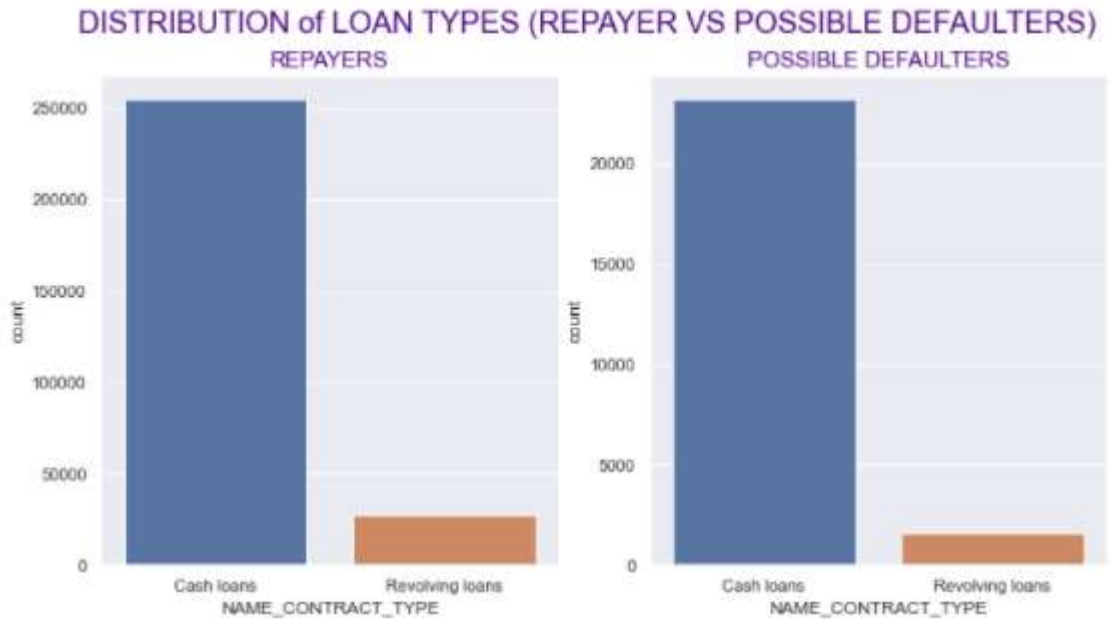
7. CLIENT AGE AND IDENTITY AGE(0.25)

8. IDENTITY AGE AND EMPLOYEMENT AGE(0.23)

9. REGISTRATION AGE AND EMPLOYEMENT AGE(0.19)

10. CLIENT PHONE AGE AND IDENTITY AGE (0.13)

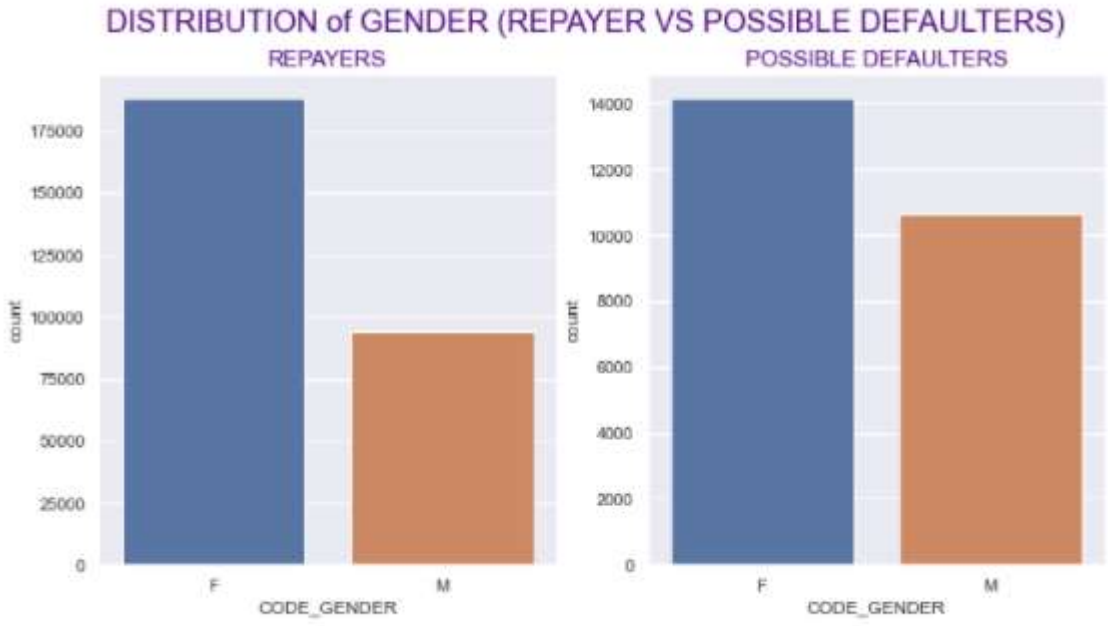
- **UNIVARIATE ANALYSIS** of categorical columns of **REPAYERS AND DEFAULTERS** data sets - Performed analysis for **NAME_CONTRACT_TYPE** as below:



OBSERVATIONS:

- There is large number of Clients taking Cash loans over Revolving loans. As seen from the plots, it can be inferred that Clients taking the Revolving loans are low risk and tend to default less.

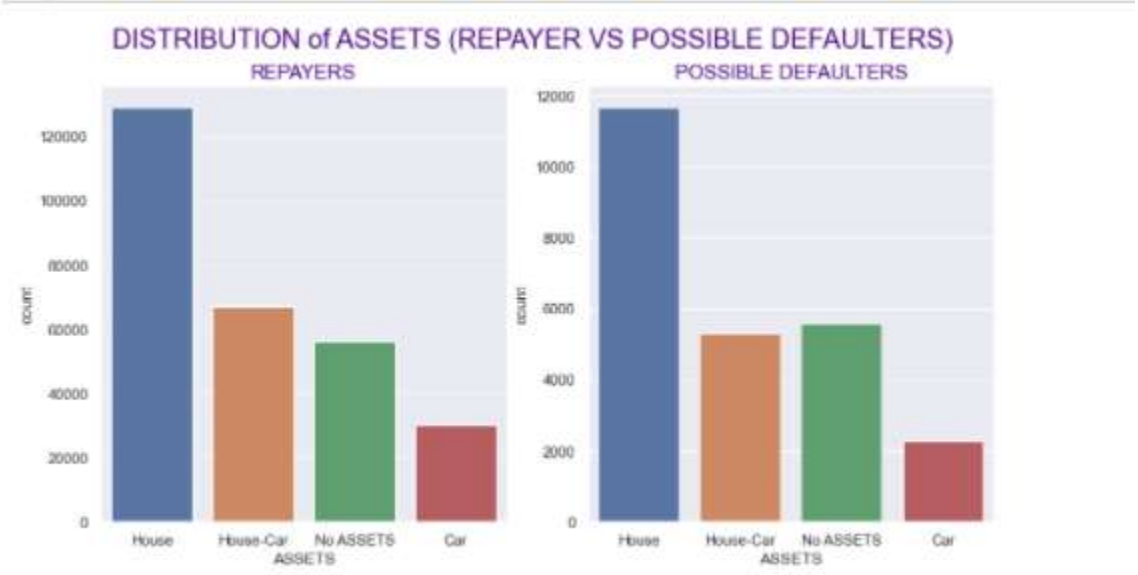
- **UNIVARIATE ANALYSIS** of categorical columns of **REPAYERS AND DEFAULTERS** data sets **-** Performed analysis for **GENDER** as below:



OBSERVATIONS:

- More number of loans are being disbursed to Females than Males. Also, the bar sizes of Males clearly show a greater Default rate when compared to Females

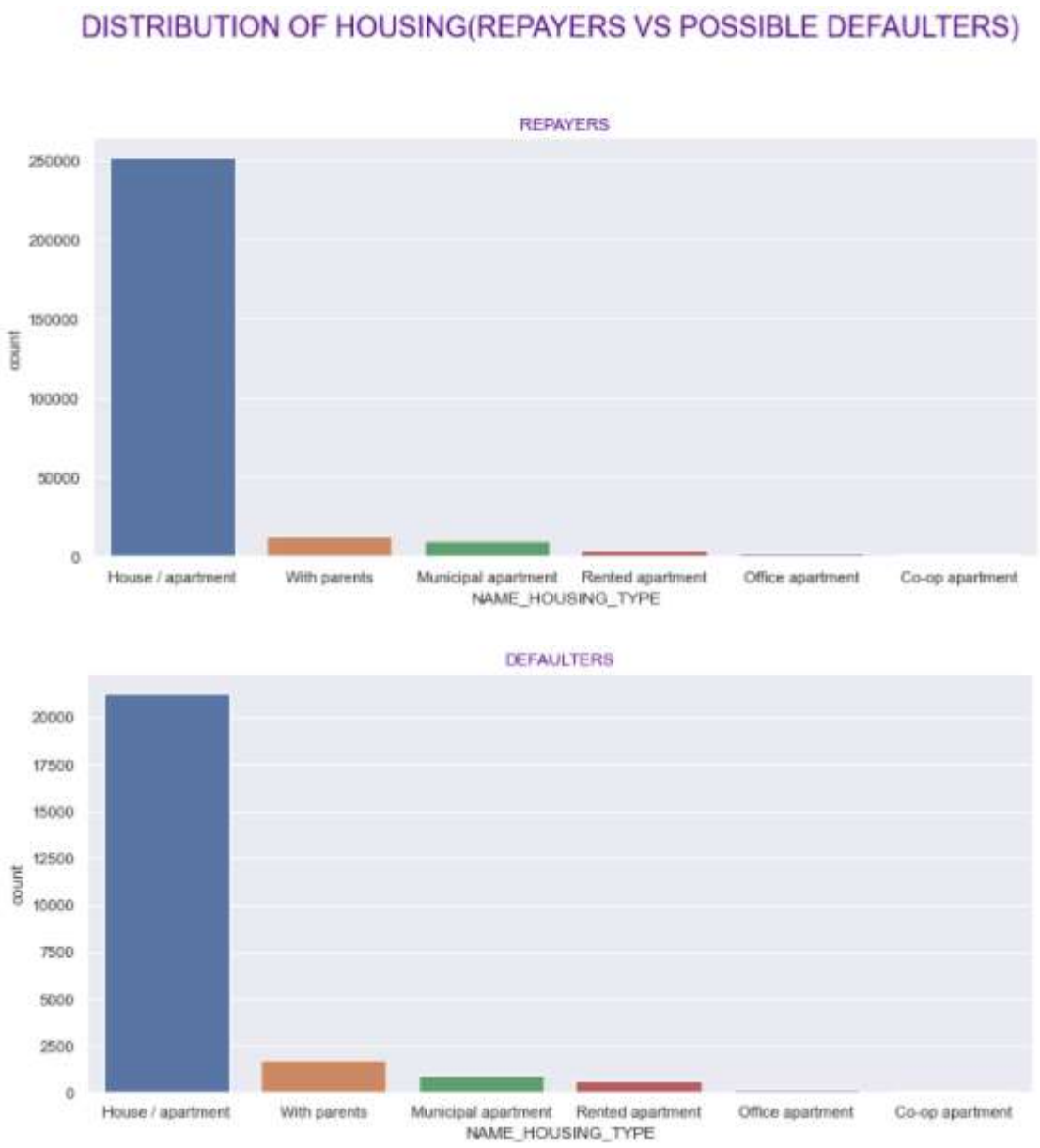
- **UNIVARIATE ANALYSIS** of categorical columns of **REPAYERS AND DEFAULTERS** data sets - Performed analysis for **ASSETS** as below:



OBSERVATIONS:

- Large number of clients seems to own a house followed by Car and House, No Assets, only Car as Asset. Also, clients with No Assets seem to be facing repaying difficulties when compared to others. Those clients who own only car seems to be low risk with relatively lesser default percentage

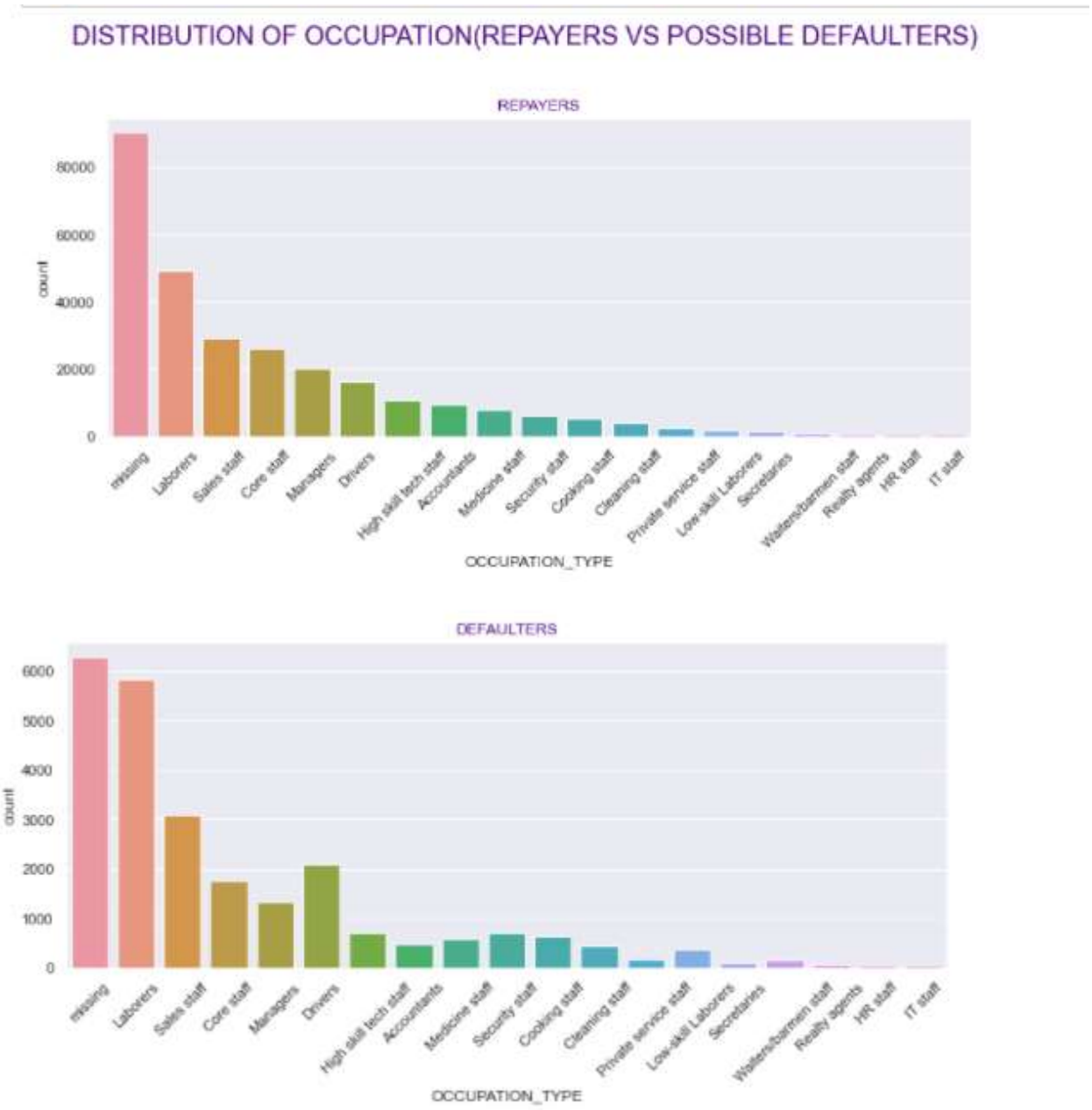
- **UNIVARIATE ANALYSIS** of categorical columns of **REPAYERS AND DEFAULTERS** data sets **-** Performed analysis for **NAME_HOUSING_TYPE** as below:



OBSERVATIONS:

- More people with House/Apartment are taking loans. As observed from bar lengths people with rented house and living with parents seem to be facing difficulty in repaying the loan when compared to other categories.

- **UNIVARIATE ANALYSIS** of categorical columns of **REPAYERS AND DEFAULTERS** data sets **-** Performed analysis for **OCCUPATION _TYPE** as below:

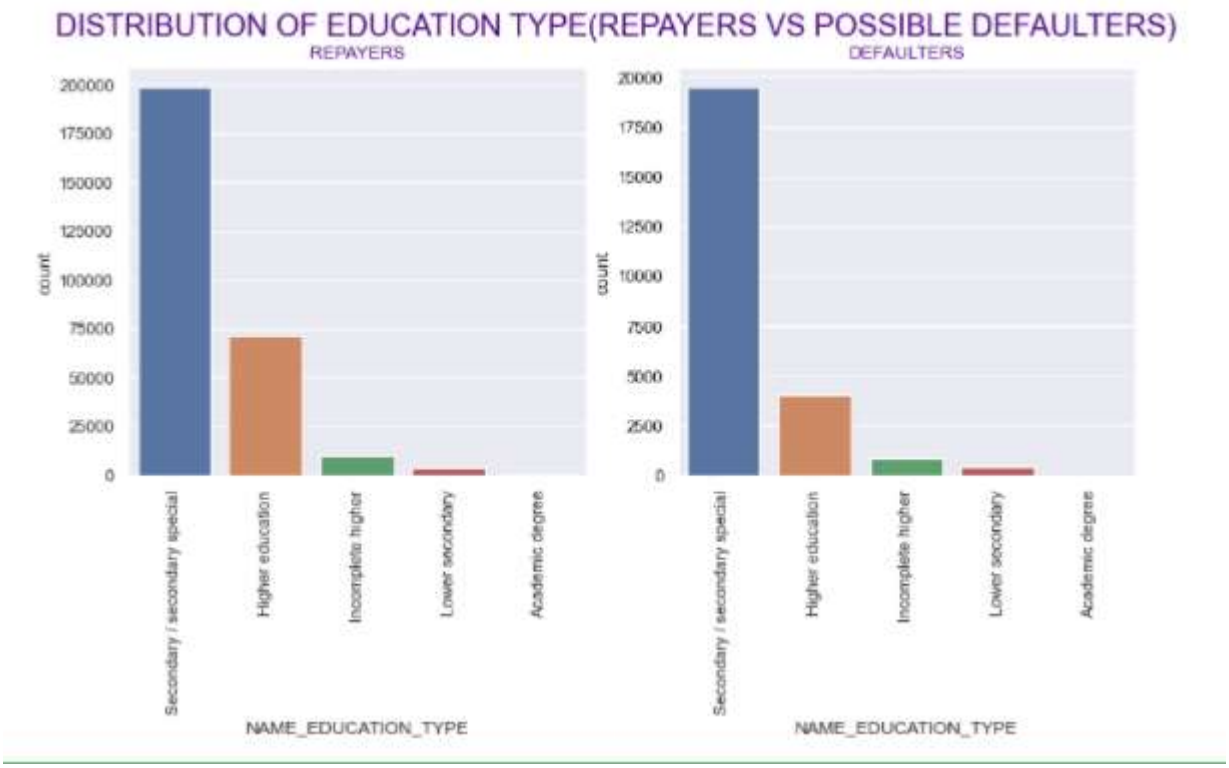


OBSERVATIONS:

- The tall bar in the missing category infers that there are many clients with occupation unknown.
- Also, the bar for laborers and drivers increased in proportion in the Defaulters plot indicating this group may be possible defaulters with paying difficulties and it is riskier to give loans to them.
- Also cooking staff, medicine staff, Security staff, cleaning staff and low skilled labourers seem to be having difficulties paying loans more when compared to others.

- The least number of clients belong to HR and IT sectors, indicating either they are not being sanctioned these loans or they are disinterested in taking Cash and Revolving Loans.
- Clients belonging to the Manager category seem to be good Repayers with low default rate.

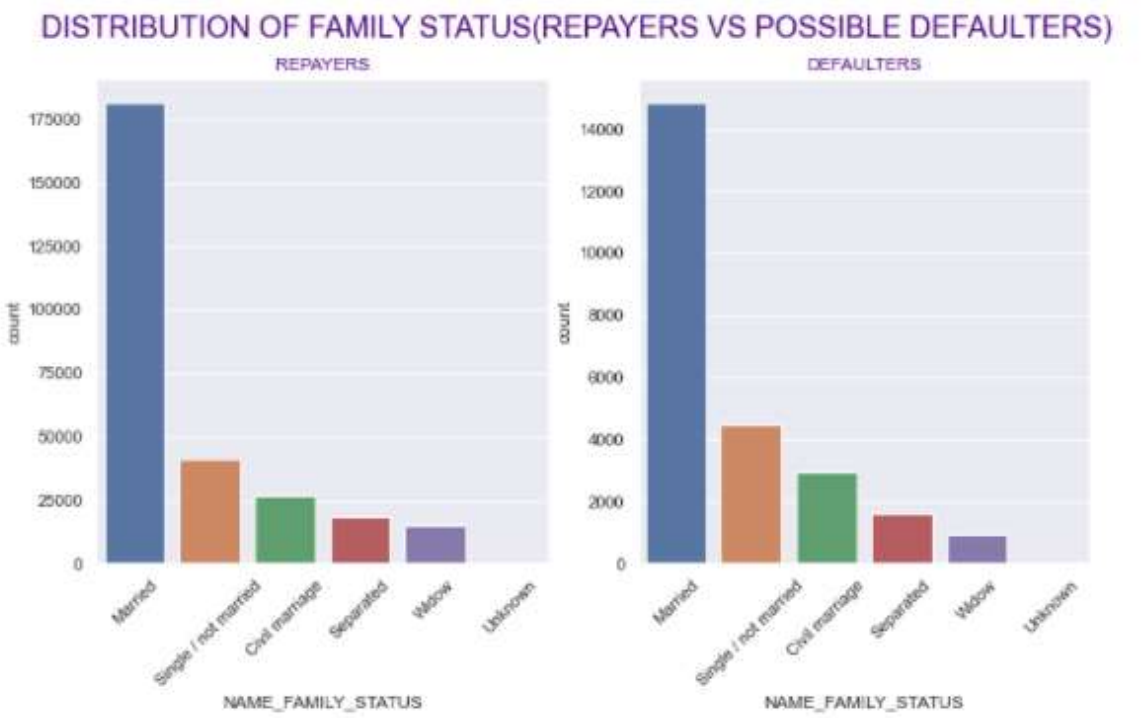
- **UNIVARIATE ANALYSIS** of categorical columns of **REPAYERS AND DEFAULTERS** data sets **-** Performed analysis for **NAME_EDUCATION_TYPE** as below:



OBSERVATIONS:

- Most of the clients seem to have persuaded Secondary/Secondary Special Education, the least of them with Academic degree.
- Clients with a Higher Education seem to be better Repayers and defaulting the loans less.

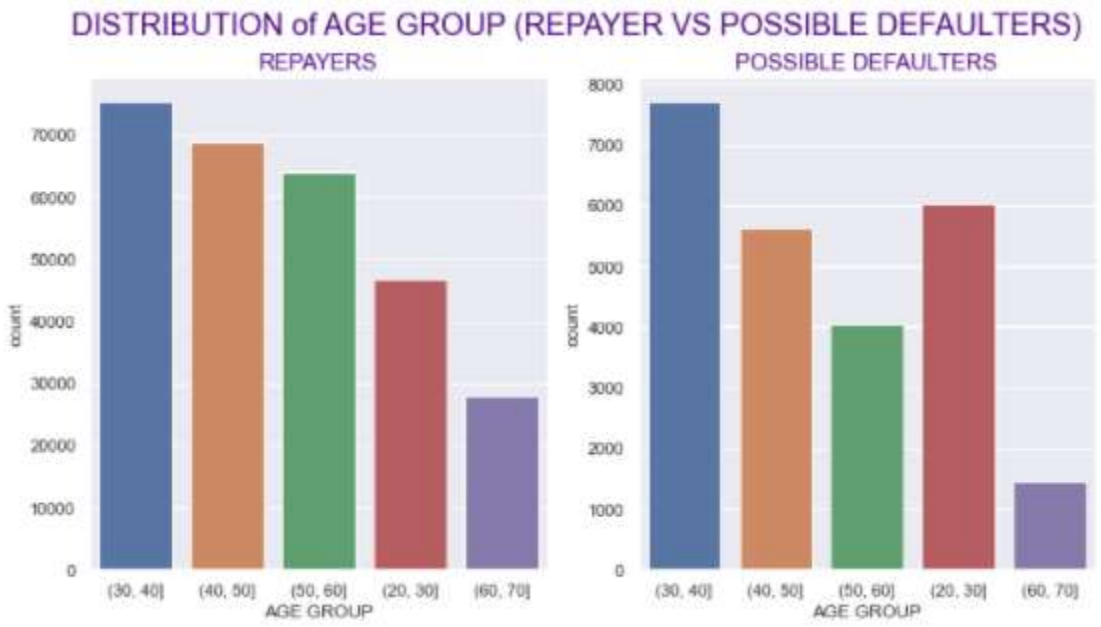
- **UNIVARIATE ANALYSIS** of categorical columns of **REPAYERS AND DEFAULTERS** data sets **-** Performed analysis for **NAME_FAMILY_STATUS** as below:



OBSERVATIONS:

- Most of the clients seem to belong to Married category in both Repayers as well as defaulters the least being Unknown followed by Widows.
- Single / not Married, Civil Marriage and Separated category people seem to have problem in paying the loans, as can be observed from a shorter bar in Repayers graph vs a longer bar in defaulters, thus indicating more percentage of them being possible Defaulters. Widows seem to have a lower default rate.
- Most of the clients Family status is known as can be inferred from a negligible at unknown category.

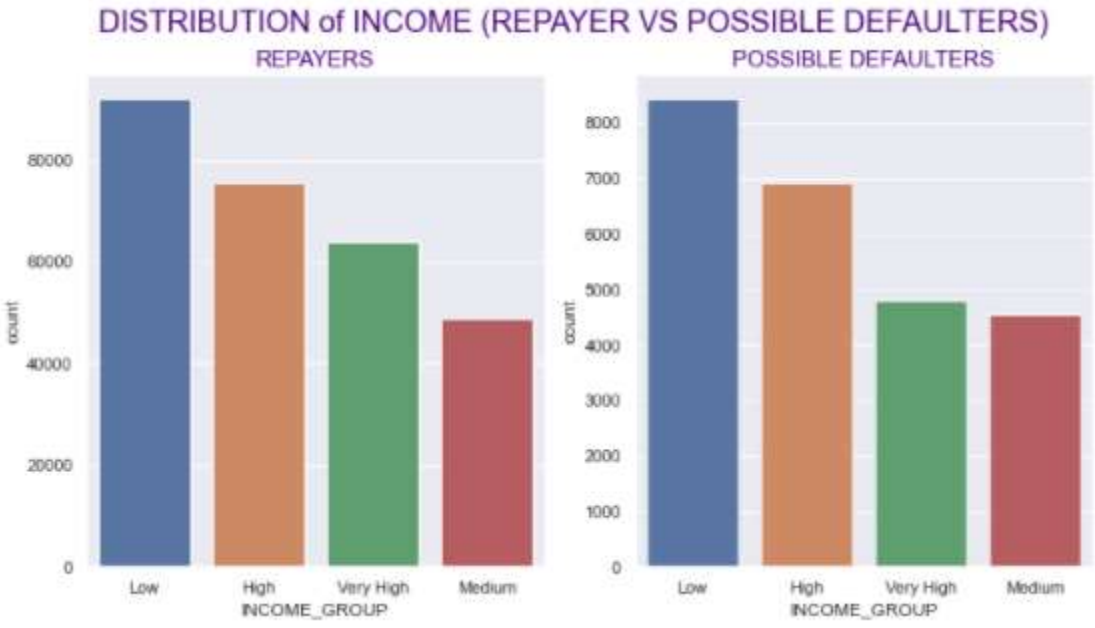
- **UNIVARIATE ANALYSIS** of categorical columns of **REPAYERS AND DEFAULTERS** data sets **-** Performed analysis for **AGE_GROUP** as below:



OBSERVATIONS:

- Most clients seem to belong to 30 to 40 age group in both Repayers as well as defaulters followed by 40 to 50 age group
- Also, the graph clearly shows a bar proportionally higher in the defaulter graph compared to Repayers graph in people aged 20 to 30 indicating a strong default rate compared to others.
- Older people belonging to 50 to 70 age group seem to be better Repayers and not defaulting on the loan.

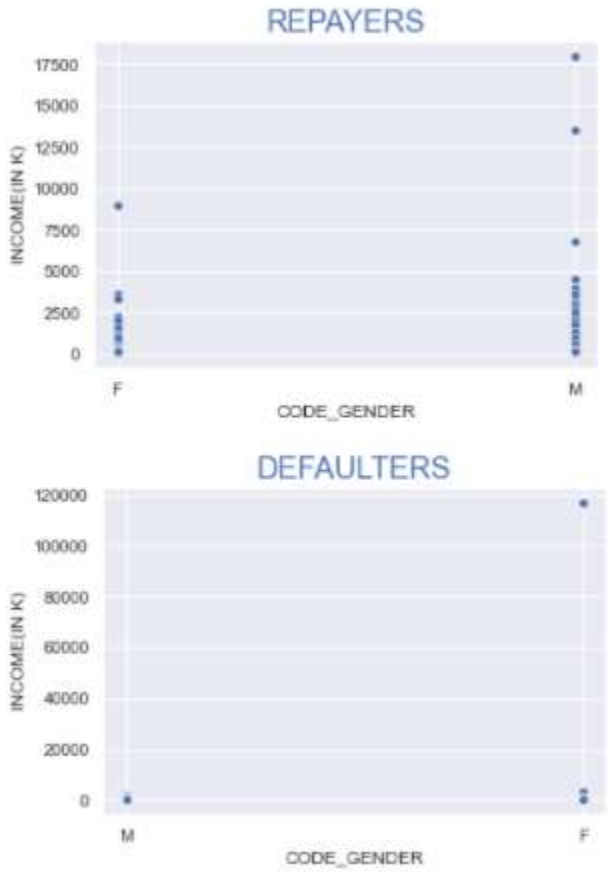
- **UNIVARIATE ANALYSIS** of categorical columns of **REPAYERS AND DEFAULTERS** data sets **-** Performed analysis for **INCOME_GROUP** as below:



OBSERVATIONS:

- Most Clients belong to Low-income binned category followed by High and Very High.
- As inferred from the graph clients with very high income are low risk customers with lesser probability of default.

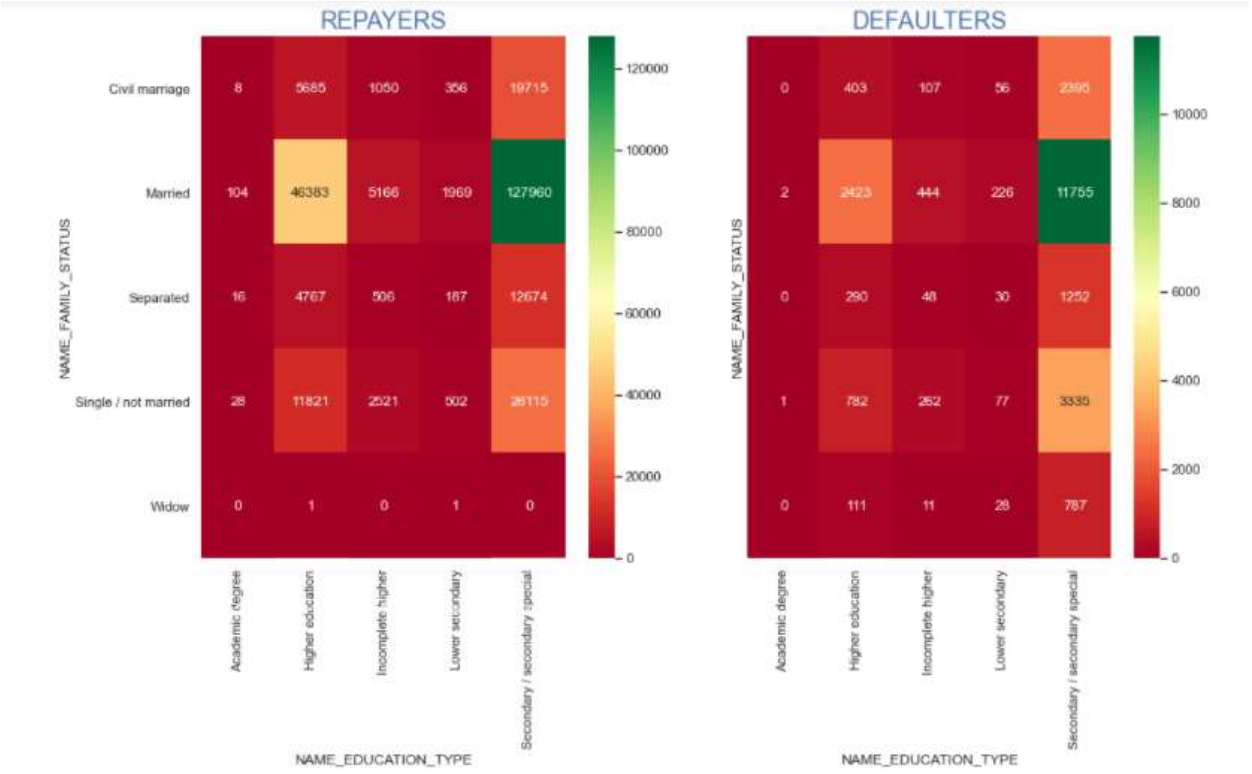
- **BI VARIATE ANALYSIS** of categorical columns of **REPAYERS AND DEFAULTERS** data sets **-** Performed analysis for **CODE_GENDER and INCOME(IN K)** as below:



OBSERVATIONS:

- It is clear and evident that Males are earning more income compared to Females in Repayers. Most Female income is segregated up to 2500K whereas males at 5000K. There are some outliers and those seem to be more for Male category.
- In Defaulters there is clearly an outlier at 117000k which needs to be further analysed. There is no big income difference observed between the two gender categories.

- **BI VARIATE ANALYSIS** of categorical columns of **REPAYERS AND DEFAULTERS data sets -** Performed analysis for **FAMILY STATUS** and **EDUCATION STATUS** as below:



OBSERVATIONS:

- Most number of married clients from both Repayers and Defaulters seems to have pursued Secondary Education followed by Higher education. Also, the family status of some clients is unknown in repayers. Most number of clients with family status Unknown belong to Higher and Lower Secondary Education.

- **STRUCTURE OF THE PREVIOUS APPLICATION DATA SET:** Carried out some inspections on the application data to check the following:

- First 5 records using head function

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKI
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	

- Number of rows and columns

```
(Rows,Columns) : (1670214, 37)
```

- Data types and missing values column wise

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1670214 entries, 0 to 1670213
Data columns (total 37 columns):
 #   Column                       Non-Null Count    Dtype  
---  ------                       --------------    -----  
 0   SK_ID_PREV                   1670214 non-null  int64  
 1   SK_ID_CURR                   1670214 non-null  int64  
 2   NAME_CONTRACT_TYPE           1670214 non-null  object 
 3   AMT_ANNUITY                  1297979 non-null  float64
 4   AMT_APPLICATION              1670214 non-null  float64
 5   AMT_CREDIT                   1670213 non-null  float64
 6   AMT_DOWN_PAYMENT             774370 non-null   float64
 7   AMT_GOODS_PRICE              1284699 non-null  float64
 8   WEEKDAY_APPR_PROCESS_START   1670214 non-null  object 
 9   HOUR_APPR_PROCESS_START      1670214 non-null  int64  
 10  FLAG_LAST_APPL_PER_CONTRACT  1670214 non-null  object 
 11  NFLAG_LAST_APPL_IN_DAY       1670214 non-null  int64  
 12  RATE_DOWN_PAYMENT            774370 non-null   float64
 13  RATE_INTEREST_PRIMARY        5951 non-null     float64
 14  RATE_INTEREST_PRIVILEGED     5951 non-null     float64
 15  NAME_CASH_LOAN_PURPOSE       1670214 non-null  object 
 16  NAME_CONTRACT_STATUS         1670214 non-null  object 
 17  DAYS_DECISION                1670214 non-null  int64  
 18  NAME_PAYMENT_TYPE            1670214 non-null  object 
 19  CODE_REJECT_REASON           1670214 non-null  object 
 20  NAME_TYPE_SUITE              849809 non-null   object 
 21  NAME_CLIENT_TYPE             1670214 non-null  object 
 22  NAME_GOODS_CATEGORY          1670214 non-null  object 
 23  NAME_PORTFOLIO               1670214 non-null  object 
 24  NAME_PRODUCT_TYPE            1670214 non-null  object 
 25  CHANNEL_TYPE                 1670214 non-null  object 
 26  SELLERPLACE_AREA             1670214 non-null  int64  
 27  NAME_SELLER_INDUSTRY         1670214 non-null  object 
 28  CNT_PAYMENT                  1297984 non-null  float64
 29  NAME_YIELD_GROUP             1670214 non-null  object 
 30  PRODUCT_COMBINATION          1669868 non-null  object 
 31  DAYS_FIRST_DRAWING           997149 non-null   float64
 32  DAYS_FIRST_DUE               997149 non-null   float64
 33  DAYS_LAST_DUE_1ST_VERSION    997149 non-null   float64
 34  DAYS_LAST_DUE                997149 non-null   float64
 35  DAYS_TERMINATION             997149 non-null   float64
 36  NFLAG_INSURED_ON_APPROVAL    997149 non-null   float64
dtypes: float64(15), int64(6), object(16)
memory usage: 471.5+ MB
```

- Null value counts and percentage of null values for each column in the data frame

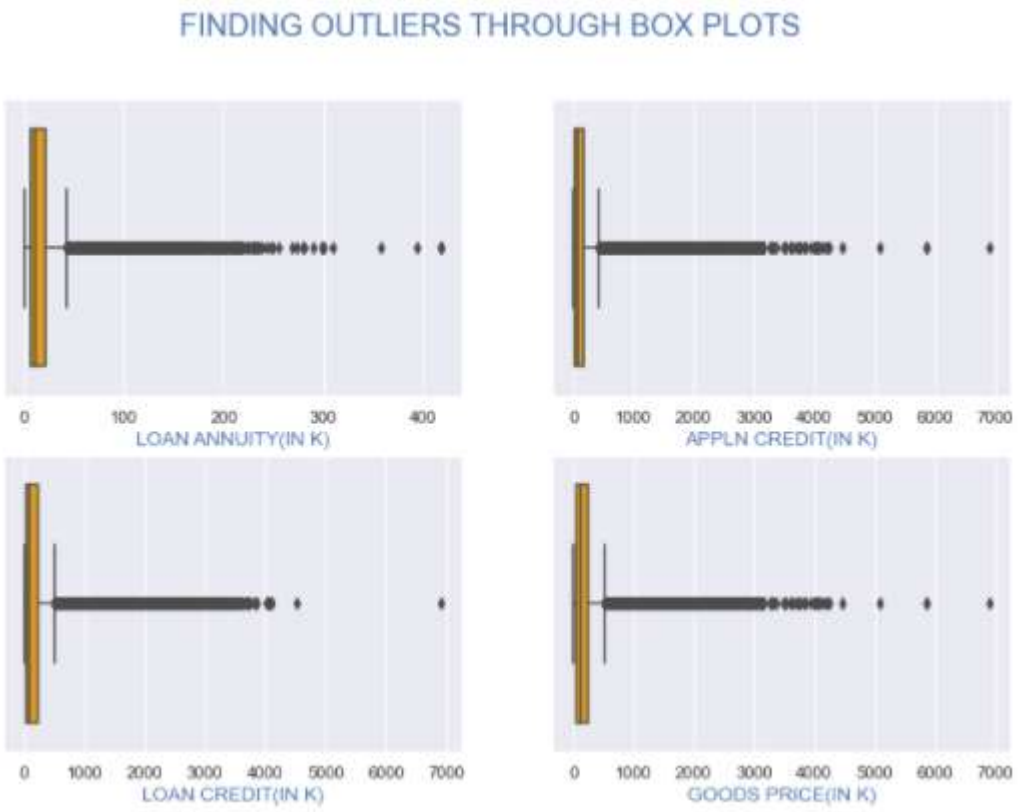
	count	Percentage
RATE_INTEREST_PRIVILEGED	1664263	99.64
RATE_INTEREST_PRIMARY	1664263	99.64
RATE_DOWN_PAYMENT	895844	53.64
AMT_DOWN_PAYMENT	895844	53.64
NAME_TYPE_SUITE	820405	49.12
NFLAG_INSURED_ON_APPROVAL	673065	40.30
DAYS_FIRST_DRAWING	673065	40.30
DAYS_FIRST_DUE	673065	40.30
DAYS_LAST_DUE_1ST_VERSION	673065	40.30
DAYS_LAST_DUE	673065	40.30
DAYS_TERMINATION	673065	40.30
AMT_GOODS_PRICE	385515	23.08
AMT_ANNUITY	372235	22.29
CNT_PAYMENT	372230	22.29
PRODUCT_COMBINATION	346	0.02
CHANNEL_TYPE	0	0.00
NAME_PRODUCT_TYPE	0	0.00
NAME_YIELD_GROUP	0	0.00
SELLERPLACE_AREA	0	0.00
NAME_SELLER_INDUSTRY	0	0.00
NAME_GOODS_CATEGORY	0	0.00
NAME_PORTFOLIO	0	0.00
SK_ID_PREV	0	0.00
NAME_CLIENT_TYPE	0	0.00
CODE_REJECT_REASON	0	0.00
SK_ID_CURR	0	0.00
DAYS_DECISION	0	0.00
NAME_CONTRACT_STATUS	0	0.00

- Created a new data frame prev_appl2 by dropping the attributes where null % > 40 , as they seem to have limited significance in the present analysis.

```
Shape of the Data Frame after dropping columns: (1670214, 26)
```

- **DATA STANDARDIZATION AND OUTLIERS:** With the new data frame, proceeded with standardizing the data and identifying the outliers.
- **AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE** - For a better visualization stored the Amount related values in thousands in new columns and dropped the originals and plotted box plots using subplots to identify the outliers.

	LOAN ANNUITY(IN K)	APPLN CREDIT(IN K)	LOAN CREDIT(IN K)	GOODS PRICE(IN K)
count	1.297979e+06	1.670214e+06	1.670213e+06	1.284699e+06
mean	1.595513e+01	1.752341e+02	1.961141e+02	2.278476e+02
std	1.478211e+01	2.927797e+02	3.185746e+02	3.153964e+02
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	6.320000e+00	1.872000e+01	2.416000e+01	5.084000e+01
50%	1.125000e+01	7.105000e+01	8.054000e+01	1.123200e+02
75%	2.066000e+01	1.803600e+02	2.164200e+02	2.340000e+02
max	4.180600e+02	6.905160e+03	6.905160e+03	6.905160e+03



OBSERVATIONS:

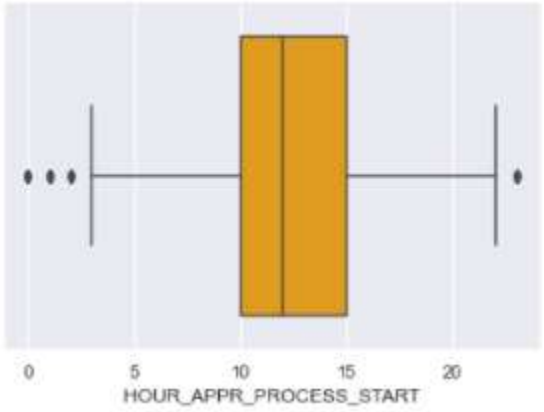
- As observed from the box plots and the summary there are outliers in LOAN_ANNUITY, APPLN CREDIT, LOAN CREDIT and GOODS PRICE. These values need to be further analysed and can be dealt with:
  1. Either by imputing them with median.
  2. Capping the columns on basis of quartiles.

- **DAYS_DECISION–** Converted the negative DAYS DECISION values to positive by using pd.abs function and converted the days to years for standardization. Then plotted box plot to identify the outliers.

OBSERVATIONS:

- As observed from the plot the records with 8 years are being treated as outliers due their high variance from the mean of actual data. These values can be imputed with median for a better analysis.

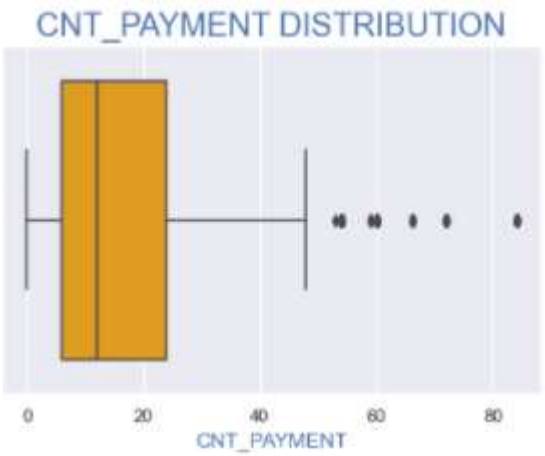
- **HOUR_APPR_PROCESS_START** - We plotted Box/count plot to check outliers.



OBSERVATIONS:

- Most of the applications are being applied between morning 10 to evening 3 (Banking working Hours), some of them are being applied online and the time does not correspond to the banking hours so they are being plotted as outliers when compared to the actual data.
- We can infer that many clients wish to apply directly through Bank rather than online.

- **CNT_PAYMENT**- We plotted box plot to check outliers.



OBSERVATION:

- Most of the data lies below 45 .There are some outliers due to high variance from the mean.

- **Duplicate/Junk value check in the object data type columns**: we checked unique values present in all the object data type columns to identify if there are any duplicates/junk values.

DUPLICATES/JUNK VALUES IN OBJECT COLUMNS

[69]:
```
#Get all the Object data type columns and thier unique values to identify if there are any duplicates or junk values
obj_cols=[col for col in prev_appl2.columns if prev_appl2[col].dtypes=='O']
for col in obj_cols:
    print(col,":\n",prev_appl2[col].unique(),"\n")
```

```
NAME_CONTRACT_TYPE :
 ['Consumer loans' 'Cash loans' 'Revolving loans' 'XNA']

WEEKDAY_APPR_PROCESS_START :
 ['SATURDAY' 'THURSDAY' 'TUESDAY' 'MONDAY' 'FRIDAY' 'SUNDAY' 'WEDNESDAY']

FLAG_LAST_APPL_PER_CONTRACT :
 ['Y' 'N']

NAME_CASH_LOAN_PURPOSE :
 ['XAP' 'XNA' 'Repairs' 'Everyday expenses' 'Car repairs'
 'Building a house or an annex' 'Other' 'Journey'
 'Purchase of electronic equipment' 'Medicine' 'Payments on other loans'
 'Urgent needs' 'Buying a used car' 'Buying a new car'
 'Buying a holiday home / land' 'Education' 'Buying a home' 'Furniture'
 'Buying a garage' 'Business development' 'Wedding / gift / holiday'
 'Hobby' 'Gasification / water supply' 'Refusal to name the goal'
 'Money for a third person']

NAME_CONTRACT_STATUS :
 ['Approved' 'Refused' 'Canceled' 'Unused offer']

NAME_PAYMENT_TYPE :
 ['Cash through the bank' 'XNA' 'Non-cash from your account'
 'Cashless from the account of the employer']

CODE_REJECT_REASON :
 ['XAP' 'HC' 'LIMIT' 'CLIENT' 'SCOFR' 'SCO' 'XNA' 'VERIF' 'SYSTEM']
```

```
NAME_CLIENT_TYPE :
 ['Repeater' 'New' 'Refreshed' 'XNA']

NAME_GOODS_CATEGORY :
 ['Mobile' 'XNA' 'Consumer Electronics' 'Construction Materials'
 'Auto Accessories' 'Photo / Cinema Equipment' 'Computers' 'Audio/Video'
 'Medicine' 'Clothing and Accessories' 'Furniture' 'Sport and Leisure'
 'Homewares' 'Gardening' 'Jewelry' 'Vehicles' 'Education'
 'Medical Supplies' 'Other' 'Direct Sales' 'Office Appliances' 'Fitness'
 'Tourism' 'Insurance' 'Additional Service' 'Weapon' 'Animals'
 'House Construction']

NAME_PORTFOLIO :
 ['POS' 'Cash' 'XNA' 'Cards' 'Cars']

NAME_PRODUCT_TYPE :
 ['XNA' 'x-sell' 'walk-in']

CHANNEL_TYPE :
 ['Country-wide' 'Contact center' 'Credit and cash offices' 'Stone'
 'Regional / Local' 'AP+ (Cash loan)' 'Channel of corporate sales'
 'Car dealer']

NAME_SELLER_INDUSTRY :
 ['Connectivity' 'XNA' 'Consumer electronics' 'Industry' 'Clothing'
 'Furniture' 'Construction' 'Jewelry' 'Auto technology' 'MLM partners'
 'Tourism']

NAME_YIELD_GROUP :
 ['middle' 'low_action' 'high' 'low_normal' 'XNA']

PRODUCT_COMBINATION :
 ['POS mobile with interest' 'Cash X-Sell: low' 'Cash X-Sell: high'
 'Cash X-Sell: middle' 'Cash Street: high' 'Cash'
 'POS household without interest' 'POS household with interest'
 'POS other with interest' 'Card X-Sell' 'POS mobile without interest'
 'Card Street' 'POS industry with interest' 'Cash Street: low'
 'POS industry without interest' 'Cash Street: middle'
 'POS others without interest' nan]
```

OBSERVATIONS: No duplicate values identified in any of the columns but there seem to be some junk values like XNA,XAP as below. All the below junk records can be dropped as they won't make any sense.

- NAME_CASH_LOAN_PURPOSE has some junk values like XNA and XAP
- NAME_PAYMENT_TYPE has a junk value XNA
- CODE_REJECT_REASON has junk values like XNA and XAP
- NAME_CLIENT_TYPE has a junk value XNA
- NAME_PRODUCT_TYPE has a junk value XNA
- NAME_SELLER_INDUSTRY has a junk value XNA
- NAME_YIELD_GROUP has a junk value XNA
- PRODUCT_COMBINATION has a junk value XNA

- **MERGING CURRENT APPLICATION DATA WITH PREVIOUS APPLICATION**: After checking the outliers for prev_appl2 data, we merged both curr_appl2 and prev_appl2 data frames using **SK_ID_CURR** and **NAME_CONTRACT_TYPE** using inner join on for further analysis.

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	NAME_TYPE_SUITE	NAME_INCO
0	100003	0	Cash loans	F	N	N	0	Family	St
1	100006	0	Cash loans	F	N	Y	0	Unaccompanied	
2	100006	0	Cash loans	F	N	Y	0	Unaccompanied	
3	100006	0	Cash loans	F	N	Y	0	Unaccompanied	
4	100006	0	Cash loans	F	N	Y	0	Unaccompanied	

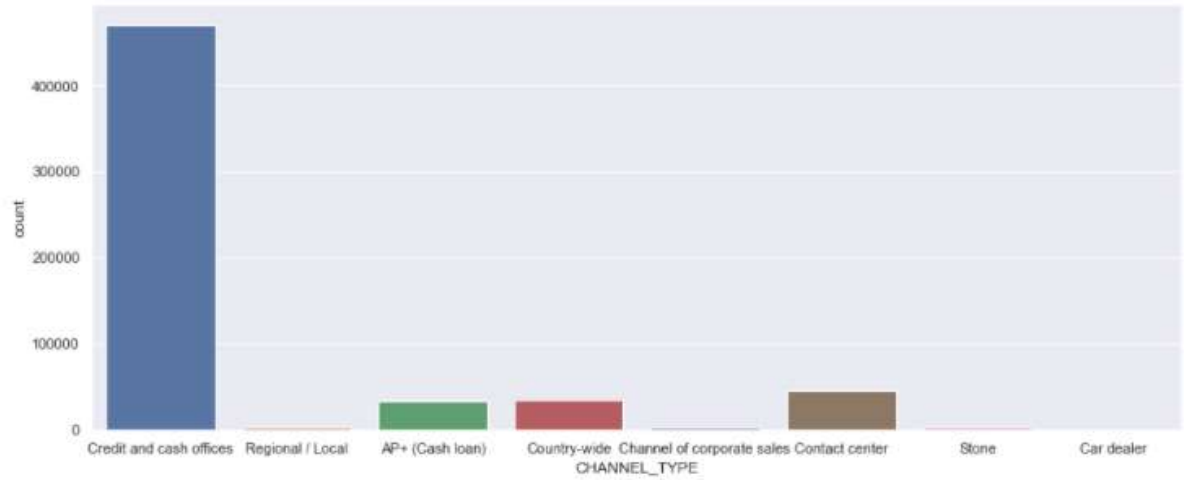
```
# Check the number of rows and columns

print("(Rows,Columns) :",curr_prev.shape)
```

```
(Rows,Columns) : (596861, 101)
```

- Checked Null Values in all the columns:

	count	Percentage
LOAN ANNUITY(IN K)_y	223286	37.41
CNT_PAYMENT	223286	37.41
GOODS PRICE(IN K)_y	222886	37.34
INCOME_GROUP	5727	0.96
SK_ID_CURR	0	0.00
AMT_REQ_CREDIT_BUREAU_YEAR	0	0.00
GOODS PRICE(IN K)_x	0	0.00
LOAN ANNUITY(IN K)_x	0	0.00
LOAN CREDIT(IN K)_x	0	0.00
INCOME(IN K)	0	0.00
CLIENT PHONE AGE(IN YRS)	0	0.00

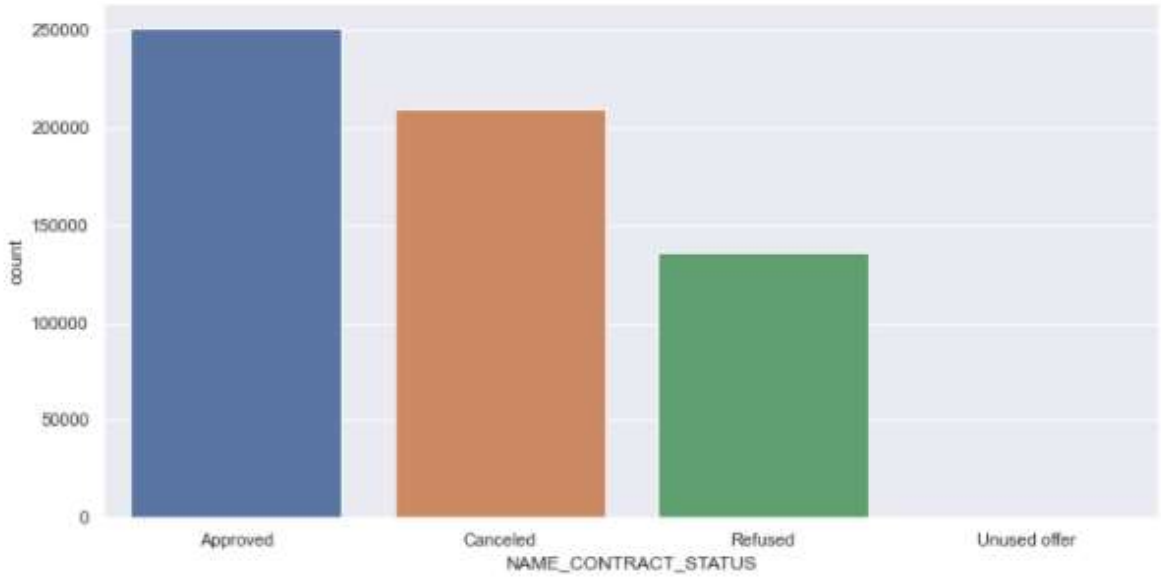
- **UNIVARIATE ANYLYSIS - CHANNEL_TYPE -** Initially we performed analysis on this categorical column by plotting as below:



OBSERVATIONS:

- Most clients are being acquired through credit and cash office followed by Contact centre. There are minimal customers being acquired through Car dealers.

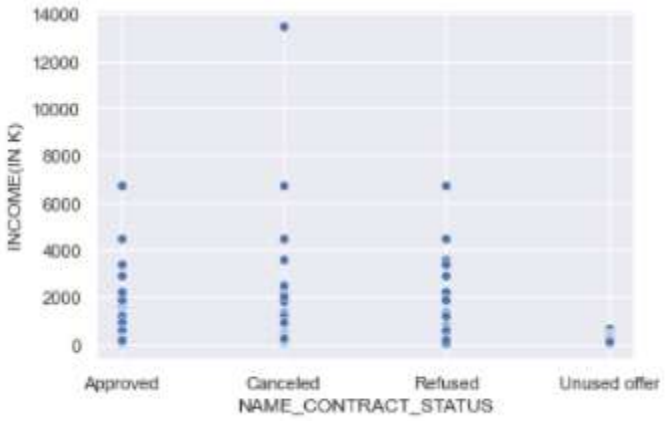
- **UNIVARIATE ANYLYSIS - NAME_CONTRACT_STATUS -** Initially we performed analysis on this categorical column by plotting as below:



OBSERVATIONS:

- Most of the loans are being approved by the bank; also many customers are cancelling the loans at some point of approval. There are very few clients who have unused the offer.

- **BIVARIATE ANYLYSIS - NAME_CONTRACT_STATUS -** Initially we performed analysis on this categorical column by plotting as below:



OBSERVATIONS:

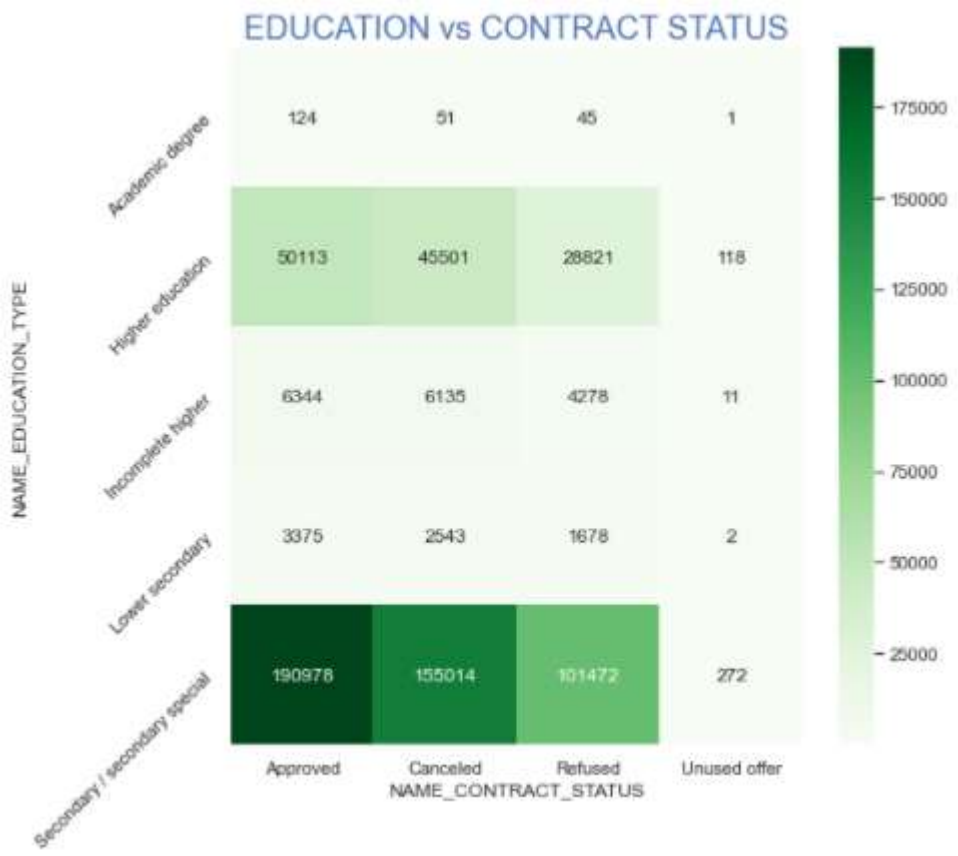
- All the unused offers are below 1000k. Income of clients is almost equally distributed among Approved, Cancelled and Refused.

- **BIVARIATE ANYLYSIS - OCCUPATION** vs **CONTRACT_STATUS -** we plotted heat map as below:

OBSERVATIONS:

- Most of the applicants from all categories of Contract type belong to Laborers followed by Sales Staff and Core Staff.

- **BIVARIATE ANYLYSIS - EDUCATION** vs **CONTRACT_STATUS -** we plotted heat map as below



OBSERVATIONS:

- Clients with Secondary Education are being approved more loans followed by Higher education. Most of the loans being cancelled belong to Secondary education group as well.

- Clients with academic degree are least in all the categories of loan contracts.

- **BIVARIATE ANYLYSIS - CHANEL** vs **CONTRACT_STATUS -** we plotted heat map as below:

OBSERVATIONS:

- Most of the loans being Approved are being channeled through credit and cash offices followed by country wide and AP+(cash loan)

END OF CASE STUDY