EDA CASE STUDY

ON CREDIT DEFAULTER

BY

RAOOF

TABLE OF CONTENT

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.’
- 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:

1. ***If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company,***
2. ***If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.***

- So using EDA we need to analyze the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Basics of EDA

EDA is an acronym for Exploratory data analysis, it refers to `Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.` - as per towardsdatascience,

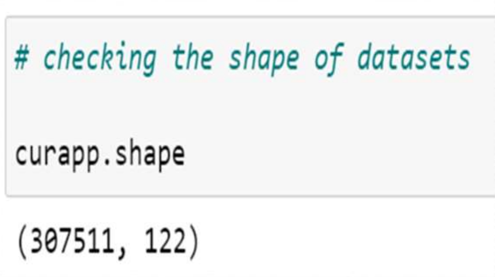
There are various steps involved in the process of EDA, vaguely they can be categorized as :

- Data reading and data understanding
- Data cleaning
- Data Analysis (both statistical and graphical)
- Drawing Inference/conclusion from our analysis

We shall look into these steps in brief further in our analysis

Data reading and Data understanding

- Data reading basically means reading the structured data (CSV in this case) into a python data frame structure.
- Once we’re done with reading the data into a data frame,
- We will analyze the shape, numerical summary, basic information and various metrics for our data frame as shown in figure below



```
# checking the shape of datasets

curapp.shape
```

```
(307511, 122)
```

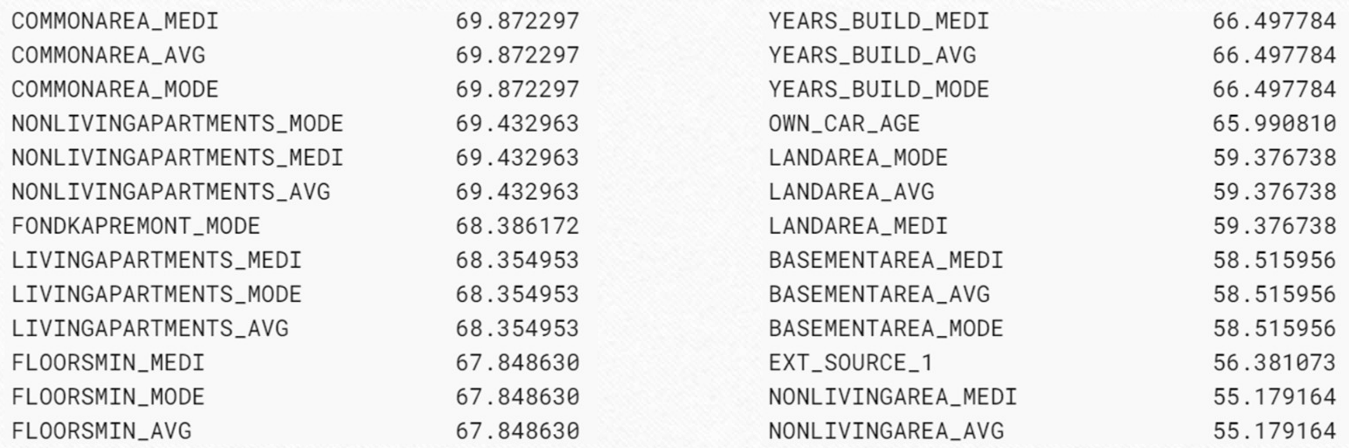
`As we can see our data frame has 307511 rows and 122 columns`

Further we can performed numerical summary and saw the basic information regarding our Data frame, during which we observed that all of the columns had correct data types associated with them, there were some unwanted columns (noise) and there were certain columns with plethora of missing values, all of these issues we shall deal with during the process of data cleaning.

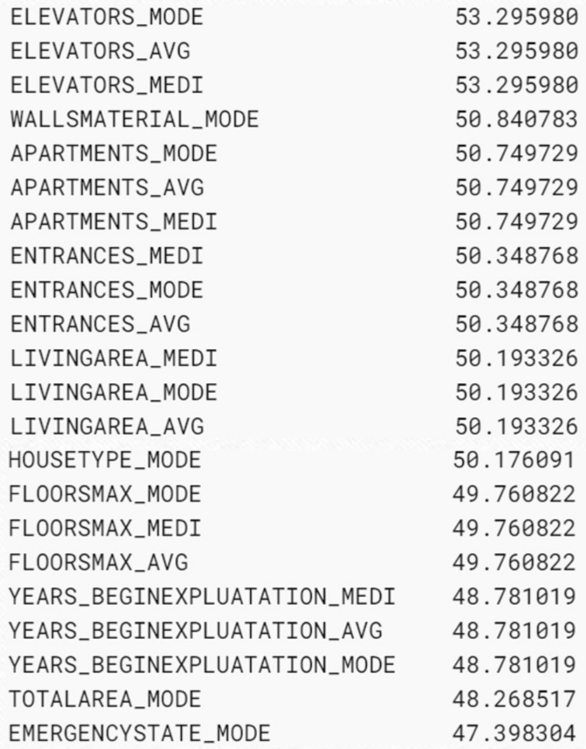
Data Cleaning

I. Dealing with Missing data A.K.A Null values

As we can see above that there are numerous columns with `larger percentage of null values` (> 40%), these columns does not contain enough information to draw insights, and in doing we so may harm our analysis by providing inadequate knowledge pertaining to significant amount of missing data, so its better to `drop them`.   Note that on left we have column name and on right their respective percentage of null values



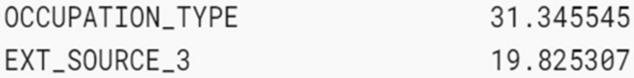
```
COMMONAREA_MEDI              69.872297        YEARS_BUILD_MEDI             66.497784
COMMONAREA_AVG               69.872297        YEARS_BUILD_AVG              66.497784
COMMONAREA_MODE              69.872297        YEARS_BUILD_MODE             66.497784
NONLIVINGAPARTMENTS_MODE     69.432963        OWN_CAR_AGE                  65.990810
NONLIVINGAPARTMENTS_MEDI     69.432963        LANDAREA_MODE                59.376738
NONLIVINGAPARTMENTS_AVG      69.432963        LANDAREA_AVG                 59.376738
FONDKAPREMONT_MODE           68.386172        LANDAREA_MEDI                59.376738
LIVINGAPARTMENTS_MEDI        68.354953        BASEMENTAREA_MEDI            58.515956
LIVINGAPARTMENTS_MODE        68.354953        BASEMENTAREA_AVG             58.515956
LIVINGAPARTMENTS_AVG         68.354953        BASEMENTAREA_MODE            58.515956
FLOORSMIN_MEDI               67.848630        EXT_SOURCE_1                 56.381073
FLOORSMIN_MODE               67.848630        NONLIVINGAREA_MEDI           55.179164
FLOORSMIN_AVG                67.848630        NONLIVINGAREA_AVG            55.179164
```



```
ELEVATORS_MODE                     53.295980
ELEVATORS_AVG                      53.295980
ELEVATORS_MEDI                     53.295980
WALLSMATERIAL_MODE                 50.840783
APARTMENTS_MODE                    50.749729
APARTMENTS_AVG                     50.749729
APARTMENTS_MEDI                    50.749729
ENTRANCES_MEDI                     50.348768
ENTRANCES_MODE                     50.348768
ENTRANCES_AVG                      50.348768
LIVINGAREA_MEDI                    50.193326
LIVINGAREA_MODE                    50.193326
LIVINGAREA_AVG                     50.193326
HOUSETYPE_MODE                     50.176091
FLOORSMAX_MODE                     49.760822
FLOORSMAX_MEDI                     49.760822
FLOORSMAX_AVG                      49.760822
YEARS_BEGINEXPLUATATION_MEDI       48.781019
YEARS_BEGINEXPLUATATION_AVG        48.781019
YEARS_BEGINEXPLUATATION_MODE       48.781019
TOTALAREA_MODE                     48.268517
EMERGENCYSTATE_MODE                47.398304
```

After dropping these columns (49 in total) we saw our shape reduced
To 307511 rows and 73 columns,

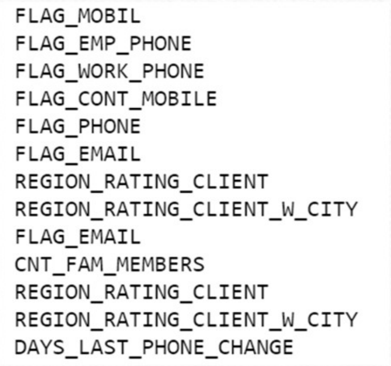
Also, We can see that the column with name `Occupation Type` has about `31% missing/null values`, we `cannot drop` that column as it still contains about 70% of relevant information, also we `cannot impute` the same as doing so may create data imbalance and bias, while there are certain techniques like logistic regression that can be used here, but for the sake of our analysis we would keep things as they are in this particular column since they won't be affecting our analysis, same stands for column `EXT_SOURCE_3`. While for all columns with missing values < 15% we simply replaced/impute them with either mode
In case of categorical column or median in case of numerical column.



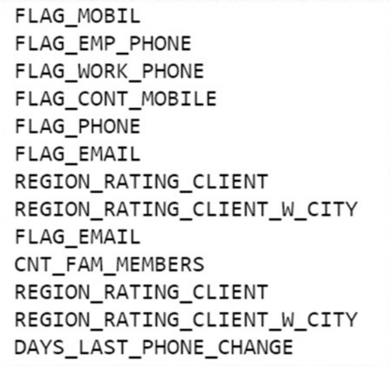
```
OCCUPATION_TYPE                 31.345545
EXT_SOURCE_3                    19.825307
```

II. Performing Sanity Checks

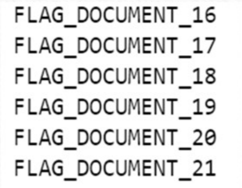
When we first looked at our data there were certain columns which needed to be removed as they offered nothing to our analysis yet made the whole process inefficient and time consuming, so we simply dropped those columns, they were namely –



```
FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
DAYS_LAST_PHONE_CHANGE
```



```
FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
DAYS_LAST_PHONE_CHANGE
```

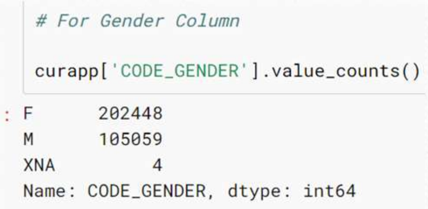


```
FLAG_DOCUMENT_16
FLAG_DOCUMENT_17
FLAG_DOCUMENT_18
FLAG_DOCUMENT_19
FLAG_DOCUMENT_20
FLAG_DOCUMENT_21
```

REMOVING UNWANTED COLUMNS

once these unwanted columns have been removed we will shift our focus on checking Value counts or simply various elements in our columns.

As we can see below that the two columns [‘Gender’] and [‘Organization type’] contain ‘XNA’ which means 'Not Available'. So we have to find the number of rows and columns and implement suitable techniques on them to fill those missing values by their corresponding modes, i.e. F in ['CODE_GENDER’] but for ['ORGANIZATION_TYPE’] we will replace them with null values, since replacing them with any other value would cause bias or data imbalance.



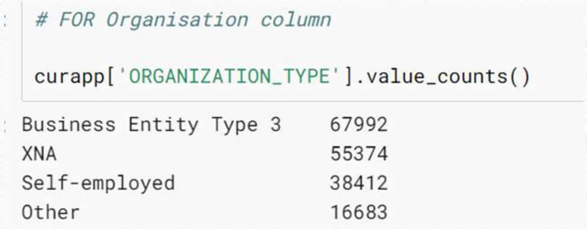
```
# For Gender Column

curapp['CODE_GENDER'].value_counts()
```

```
F      202448
M      105059
XNA         4
Name: CODE_GENDER, dtype: int64
```

CHECKING EELEMNTS OF OUR COLUMN

Note that [‘ORGANIZATION TYPE’] column have more members but only top 4 were shown in this presentation for readability purposes

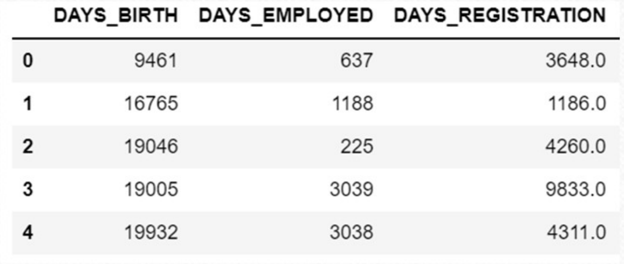


```
# FOR Organisation column

curapp['ORGANIZATION_TYPE'].value_counts()
```

```
Business Entity Type 3     67992
XNA                        55374
Self-employed              38412
Other                      16683
```

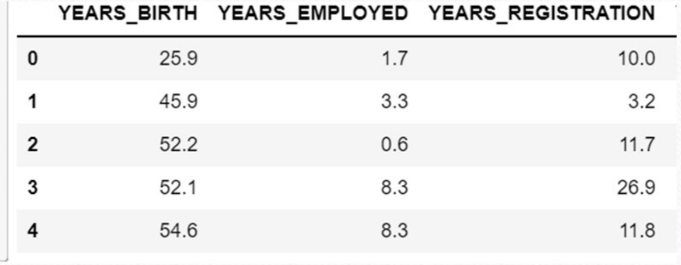
when we look at the columns ['DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH'] which contains Days in them we noticed some negative values which shouldn't be there as time can't be negative so simply replaced all their values by their absolute form and converted the same into years for better readability, also we noticed that there was simply no need to change the format of these particulars, in fact all of our columns practically had the correct data type associated to them,



	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION
0	9461	637	3648.0
1	16765	1188	1186.0
2	19046	225	4260.0
3	19005	3039	9833.0
4	19932	3038	4311.0

← *BEFORE TRANFOSRMATION*

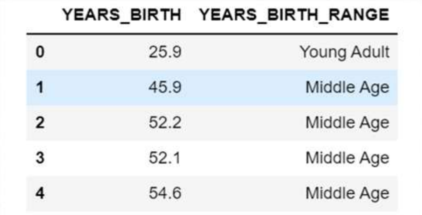
AFTER TRANSFORMATION →



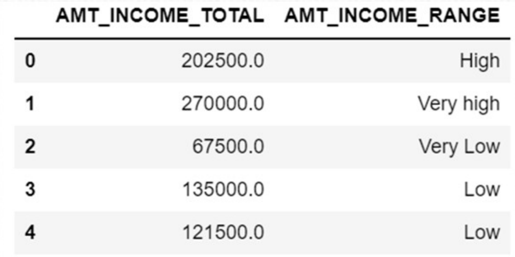
	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_REGISTRATION
0	25.9	1.7	10.0
1	45.9	3.3	3.2
2	52.2	0.6	11.7
3	52.1	8.3	26.9
4	54.6	8.3	11.8

-- BINNING OF CONTINOUS VARIABLES --

Many columns like the ones containing 'Years' and 'Amount' associated values in them, though seemed discrete but in fact were continuous in nature so it only made sense to Bin them or bucket them for better analysis -:



	YEARS_BIRTH	YEARS_BIRTH_RANGE
0	25.9	Young Adult
1	45.9	Middle Age
2	52.2	Middle Age
3	52.1	Middle Age
4	54.6	Middle Age

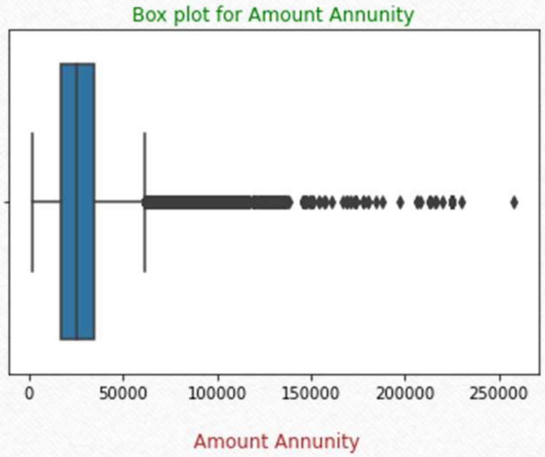


	AMT_INCOME_TOTAL	AMT_INCOME_RANGE
0	202500.0	High
1	270000.0	Very high
2	67500.0	Very Low
3	135000.0	Low
4	121500.0	Low

UNIVARIATE DATA ANALYSIS

-- Box Plot --

- We begin with univariate analysis so as to check the data for outliers if any, then analyze the data for any trends or patterns.



One of the best way to achieve the above objective is by using a boxplot, it basically graphically depicts our numerical data using quartiles, any values present outside of it's IQR or top fence is treated as an outliers, as we can see for the same in the figure to the left,

If we observe the boxplot for Amount Annuity, we can see that - :

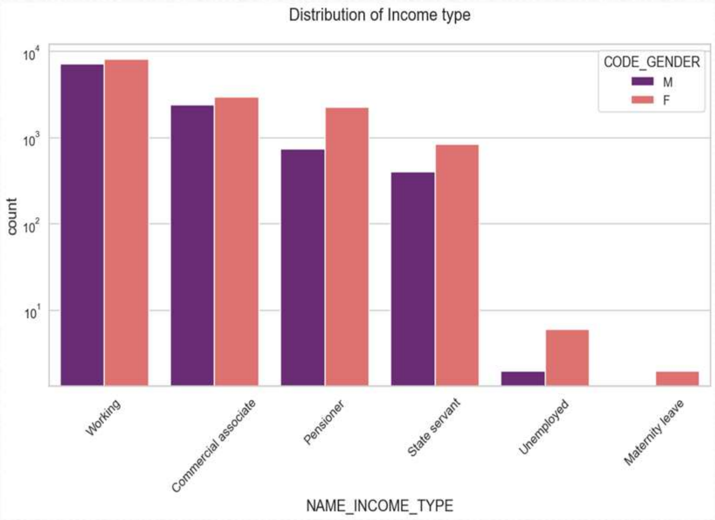
The median or 50th percentile lies somewhere around 30,000 while IQR ends at about 65,000 and anything beyond that is deemed as an outlier,

From above we can conclude that though there are outliers present in this column these values are continuous and nothing out of ordinary, but If we want to we can impute some value of central tendency in place of these outliers so as to eliminate them or we can simply bin/cap the data in our column.

UNIVARIATE DATA ANALYSIS

-- Count Plot --

Using count plot we can see whether there is any specific pattern in our data, We Can see from the count plot of 'Distribution of Income' The basic trend that our data follows and from which we can conclude that, :

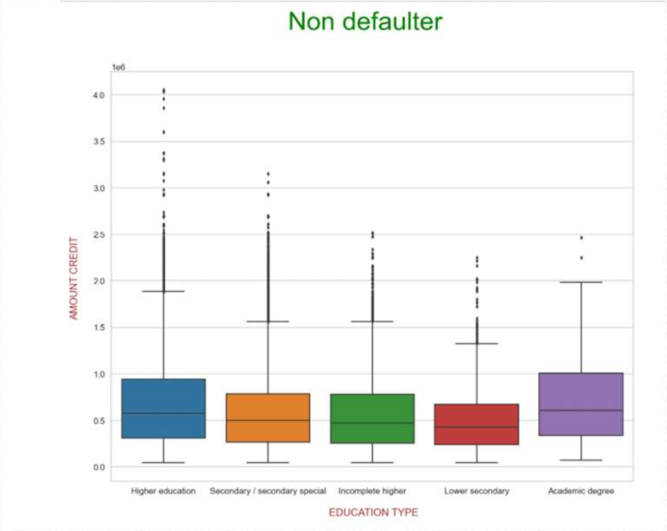


1. For income type 'working', 'commercial associate' and 'State Servant' the number of credits are higher than other i.e. 'Maternity leave'.
2. For this Females are having more number of credits than male.
3. Less number of credits for income type 'Maternity leave'.
4. For type 1: There is no income type for 'student', 'pensioner' an 'Businessman' which means they don't do any late payments.

BIVARIATE DATA ANALYSIS

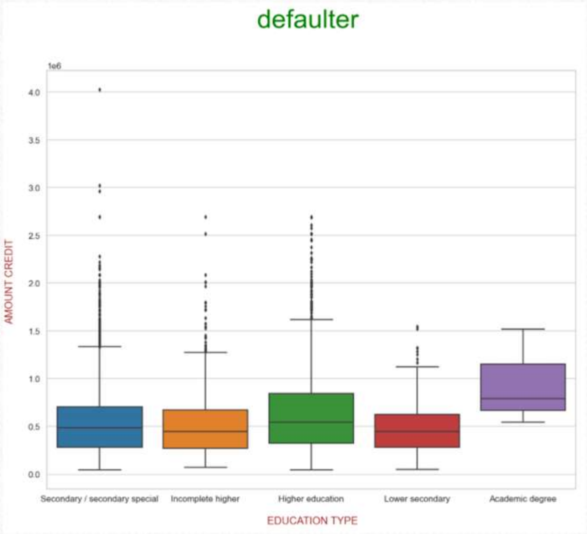
-- Box Plot --

- We can compare two metrics together using a box plot [Categorical and Numerical], here though we would be comparing Education type and Amount of credit requested by the client, for Non-Defaulter and Defaulter,



Here we can notice that for both defaulters and non-defaulters, the more educated they are more credit they ask for,

While it can also be seen that no defaulter with academic degree has ever asked for less than 500,000 in credit

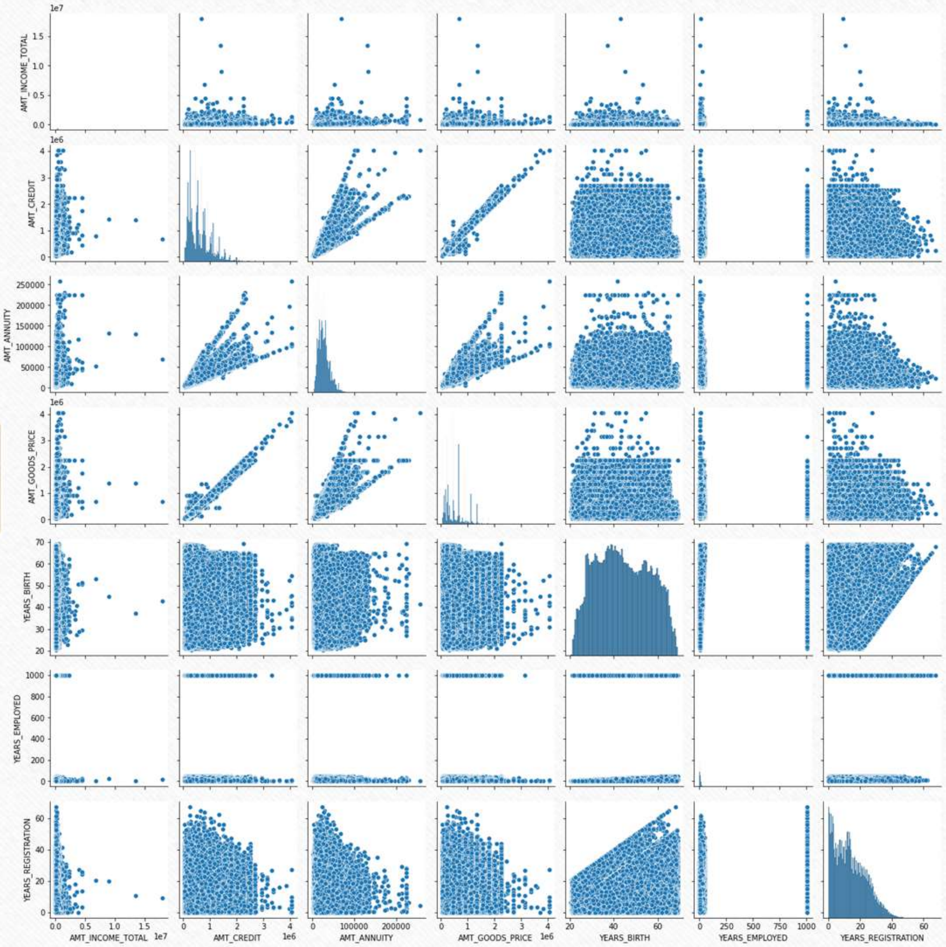


BIVARIATE DATA ANALYSIS

-- Pair Plot --

- We can compare two metrics together using a pair plot [both Numerical], here though we would be comparing all metrics between columns pertaining to Amount as basis, and we will do so for both Non-Defaulters and Defaulters separately.

- The Pair Plot basically, plots a matrix of relationships between various entities which are numerical in nature, such that all diagonal elements are Histograms or distribution plot of itself, while all other elements are shown by scatter plot between the respective row and column indices of that position.



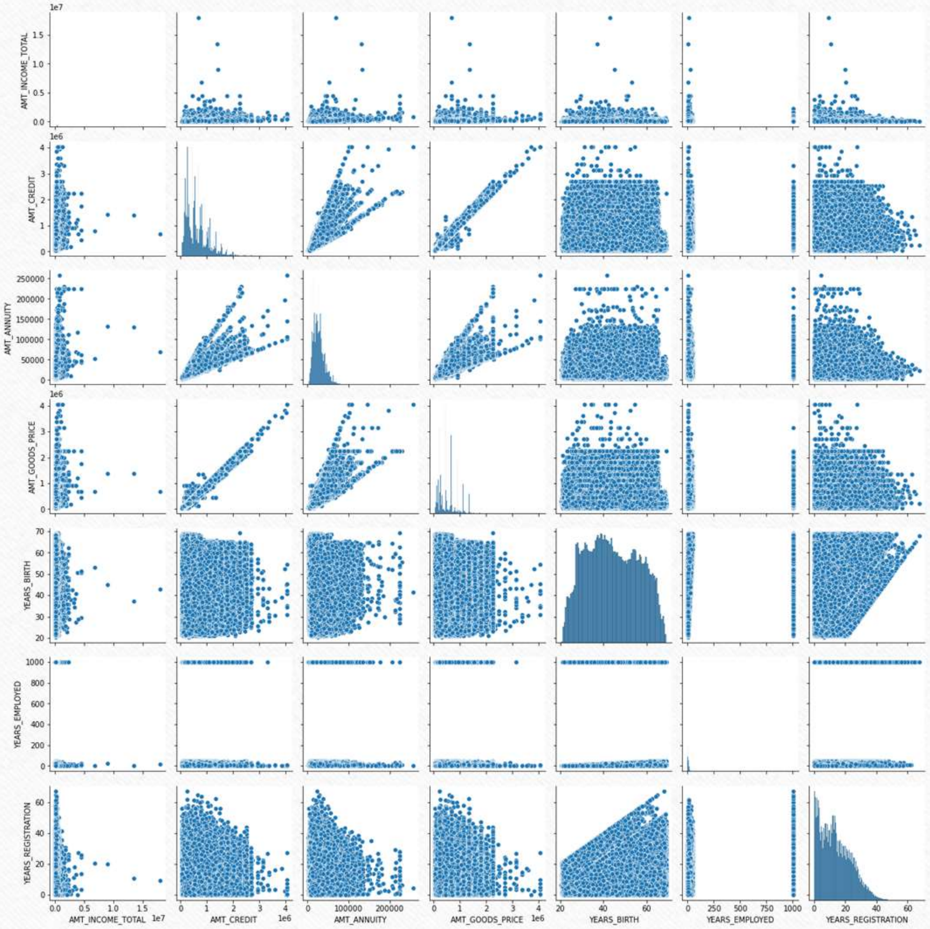
Pair Plot for Non-Defaulters

Here we formed a pair plot between all columns with metrics related to amount for Non-Defaulters,

From the plot it is evident that top correlations between these metrics are as follows - :

- Amount of credit and Amount of goods price,
- Amount of credit and Amount of Annuity,
- Amount of Annuity and Amount of goods price

While Years Birth and Years Registration do have some correlation but nothing too significant.



Pair Plot for Defaulters

Here we formed a pair plot between all columns with metrics related to amount for Defaulters,

From the plot it is evident that top correlations are same as in case of Non- Defaulters i.e., they are - :

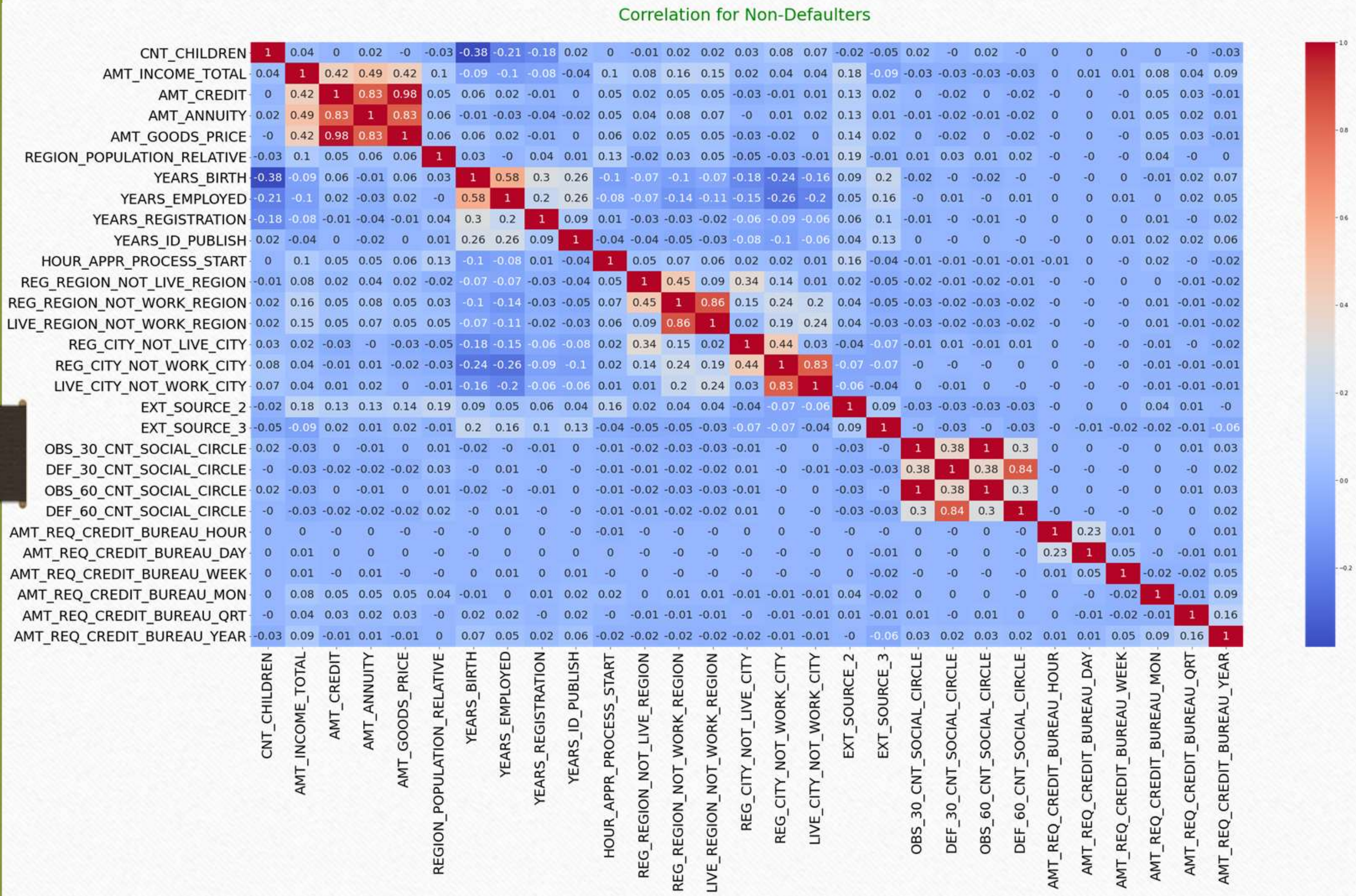
- Amount of credit and Amount of goods price,
- Amount of credit and Amount of Annuity,
- Amount of Annuity and Amount of goods price

Similarly as before Years Birth and Years Registration do have some correlation but nothing too significant.

BIVARIATE DATA ANALYSIS

-- Heat Maps--

- Using Heat Maps we can see compare any two numerical data, or even plethora of categorical data's given there must be some numerical column present so as to establish correlation between those entities, We will draw a heatmap between all the numerical data present in our dataset, for both Non-Defaulters and Defaulters,
- Similarly to pair plots, heatmaps also forms a matrix of relationships between various entities, such that all diagonal elements are Histograms or distribution plot of itself, while all other elements are shown by Correlation between the respective row and column indices of that position.

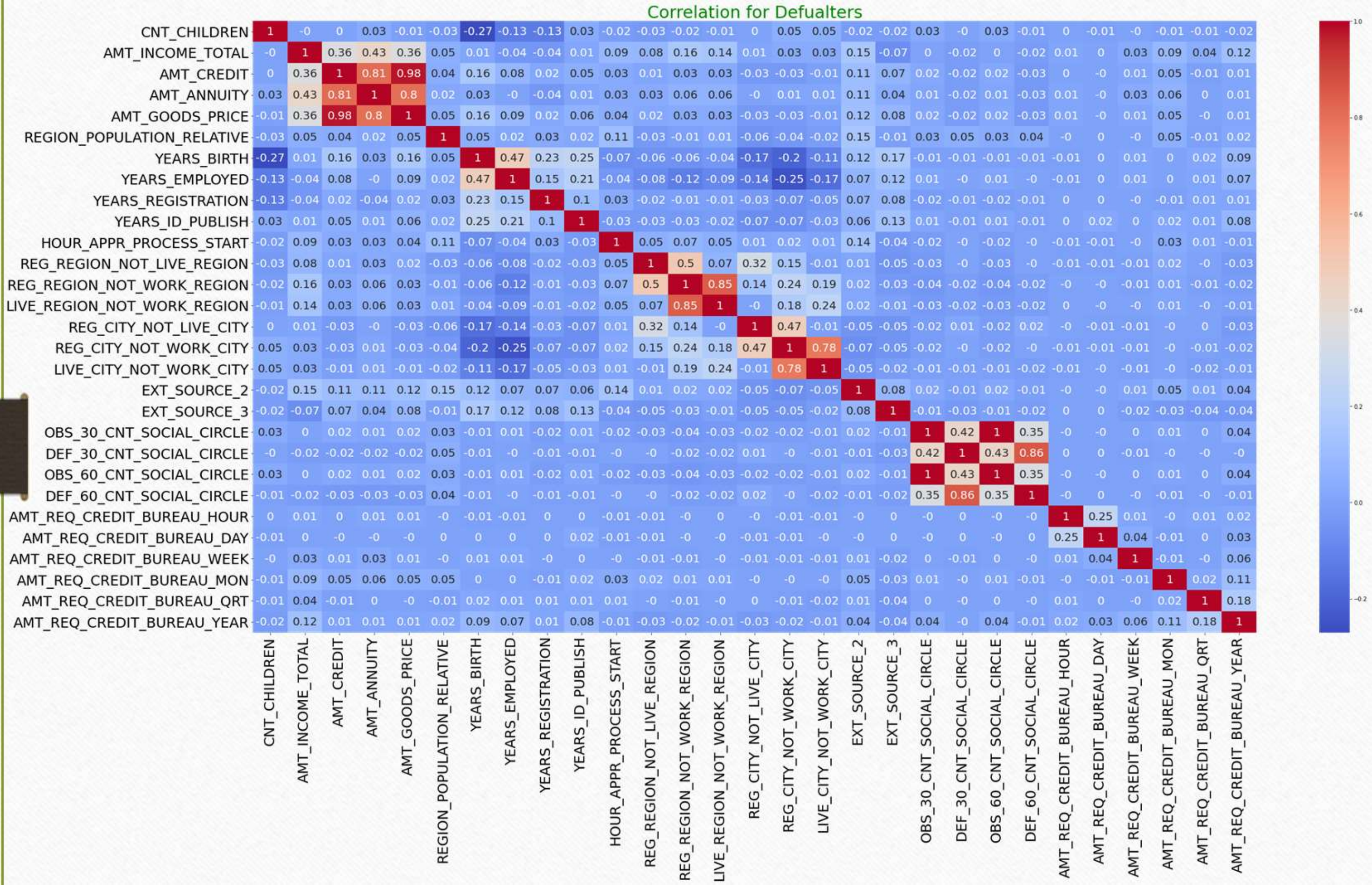


Correlation for Non-Defaulters

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_REGISTRATION	YEARS_ID_PUBLISH	HOUR_APPR_PROCESS_START	REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	REG_CITY_NOT_LIVE_CITY	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	EXT_SOURCE_2	EXT_SOURCE_3	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	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	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1	0.04	0	0.02	-0	-0.03	-0.38	-0.21	-0.18	0.02	0	-0.01	0.02	0.02	0.03	0.08	0.07	-0.02	-0.05	0.02	-0	0.02	-0	0	0	0	0	-0	-0.03
AMT_INCOME_TOTAL	0.04	1	0.42	0.49	0.42	0.1	-0.09	-0.1	-0.08	-0.04	0.1	0.08	0.16	0.15	0.02	0.04	0.04	0.18	-0.09	-0.03	-0.03	-0.03	-0.03	0	0.01	0.01	0.08	0.04	0.09
AMT_CREDIT	0	0.42	1	0.83	0.98	0.05	0.06	0.02	-0.01	0	0.05	0.02	0.05	0.05	-0.03	-0.01	0.01	0.13	0.02	0	-0.02	0	-0.02	-0	0	-0	0.05	0.03	-0.01
AMT_ANNUITY	0.02	0.49	0.83	1	0.83	0.06	-0.01	-0.03	-0.04	-0.02	0.05	0.04	0.08	0.07	-0	0.01	0.02	0.13	0.01	-0.01	-0.02	-0.01	-0.02	0	0	0.01	0.05	0.02	0.01
AMT_GOODS_PRICE	-0	0.42	0.98	0.83	1	0.06	0.06	0.02	-0.01	0	0.06	0.02	0.05	0.05	-0.03	-0.02	0	0.14	0.02	0	-0.02	0	-0.02	-0	0	-0	0.05	0.03	-0.01
REGION_POPULATION_RELATIVE	-0.03	0.1	0.05	0.06	0.06	1	0.03	-0	0.04	0.01	0.13	-0.02	0.03	0.05	-0.05	-0.03	-0.01	0.19	-0.01	0.01	0.03	0.01	0.02	-0	-0	-0	0.04	-0	0
YEARS_BIRTH	-0.38	-0.09	0.06	-0.01	0.06	0.03	1	0.58	0.3	0.26	-0.1	-0.07	-0.1	-0.07	-0.18	-0.24	-0.16	0.09	0.2	-0.02	-0	-0.02	-0	-0	-0	0	-0.01	0.02	0.07
YEARS_EMPLOYED	-0.21	-0.1	0.02	-0.03	0.02	-0	0.58	1	0.2	0.26	-0.08	-0.07	-0.14	-0.11	-0.15	-0.26	-0.2	0.05	0.16	-0	0.01	-0	0.01	0	0	0.01	0	0.02	0.05
YEARS_REGISTRATION	-0.18	-0.08	-0.01	-0.04	-0.01	0.04	0.3	0.2	1	0.09	0.01	-0.03	-0.03	-0.02	-0.06	-0.09	-0.06	0.06	0.1	-0.01	-0	-0.01	-0	0	0	0	0.01	-0	0.02
YEARS_ID_PUBLISH	0.02	-0.04	0	-0.02	0	0.01	0.26	0.26	0.09	1	-0.04	-0.04	-0.05	-0.03	-0.08	-0.1	-0.06	0.04	0.13	0	-0	0	-0	-0	0	0.01	0.02	0.02	0.06
HOUR_APPR_PROCESS_START	0	0.1	0.05	0.05	0.06	0.13	-0.1	-0.08	0.01	-0.04	1	0.05	0.07	0.06	0.02	0.02	0.01	0.16	-0.04	-0.01	-0.01	-0.01	-0.01	-0.01	0	-0	0.02	-0	-0.02
REG_REGION_NOT_LIVE_REGION	-0.01	0.08	0.02	0.04	0.02	-0.02	-0.07	-0.07	-0.03	-0.04	0.05	1	0.45	0.09	0.34	0.14	0.01	0.02	-0.05	-0.02	-0.01	-0.02	-0.01	-0	-0	0	0	-0.01	-0.02
REG_REGION_NOT_WORK_REGION	0.02	0.16	0.05	0.08	0.05	0.03	-0.1	-0.14	-0.03	-0.05	0.07	0.45	1	0.86	0.15	0.24	0.2	0.04	-0.05	-0.03	-0.02	-0.03	-0.02	-0	-0	-0	0.01	-0.01	-0.02
LIVE_REGION_NOT_WORK_REGION	0.02	0.15	0.05	0.07	0.05	0.05	-0.07	-0.11	-0.02	-0.03	0.06	0.09	0.86	1	0.02	0.19	0.24	0.04	-0.03	-0.03	-0.02	-0.03	-0.02	-0	-0	-0	0.01	-0.01	-0.02
REG_CITY_NOT_LIVE_CITY	0.03	0.02	-0.03	-0	-0.03	-0.05	-0.18	-0.15	-0.06	-0.08	0.02	0.34	0.15	0.02	1	0.44	0.03	-0.04	-0.07	-0.01	0.01	-0.01	0.01	0	-0	-0	-0.01	-0	-0.02
REG_CITY_NOT_WORK_CITY	0.08	0.04	-0.01	0.01	-0.02	-0.03	-0.24	-0.26	-0.09	-0.1	0.02	0.14	0.24	0.19	0.44	1	0.83	-0.07	-0.07	-0	-0	-0	0	0	-0	-0	-0.01	-0.01	-0.01
LIVE_CITY_NOT_WORK_CITY	0.07	0.04	0.01	0.02	0	-0.01	-0.16	-0.2	-0.06	-0.06	0.01	0.01	0.2	0.24	0.03	0.83	1	-0.06	-0.04	0	-0.01	0	-0	-0	-0	-0	-0.01	-0.01	-0.01
EXT_SOURCE_2	-0.02	0.18	0.13	0.13	0.14	0.19	0.09	0.05	0.06	0.04	0.16	0.02	0.04	0.04	-0.04	-0.07	-0.06	1	0.09	-0.03	-0.03	-0.03	-0.03	-0	0	0	0.04	0.01	-0
EXT_SOURCE_3	-0.05	-0.09	0.02	0.01	0.02	-0.01	0.2	0.16	0.1	0.13	-0.04	-0.05	-0.05	-0.03	-0.07	-0.07	-0.04	0.09	1	-0	-0.03	-0	-0.03	-0	-0.01	-0.02	-0.02	-0.01	-0.06
OBS_30_CNT_SOCIAL_CIRCLE	0.02	-0.03	0	-0.01	0	0.01	-0.02	-0	-0.01	0	-0.01	-0.02	-0.03	-0.03	-0.01	-0	0	-0.03	-0	1	0.38	1	0.3	0	0	-0	0	0.01	0.03
DEF_30_CNT_SOCIAL_CIRCLE	-0	-0.03	-0.02	-0.02	-0.02	0.03	-0	0.01	-0	-0	-0.01	-0.01	-0.02	-0.02	0.01	-0	-0.01	-0.03	-0.03	0.38	1	0.38	0.84	-0	-0	-0	0	-0	0.02
OBS_60_CNT_SOCIAL_CIRCLE	0.02	-0.03	0	-0.01	0	0.01	-0.02	-0	-0.01	0	-0.01	-0.02	-0.03	-0.03	-0.01	-0	0	-0.03	-0	1	0.38	1	0.3	0	0	-0	0	0.01	0.03
DEF_60_CNT_SOCIAL_CIRCLE	-0	-0.03	-0.02	-0.02	-0.02	0.02	-0	0.01	-0	-0	-0.01	-0.01	-0.02	-0.02	0.01	0	-0	-0.03	-0.03	0.3	0.84	0.3	1	-0	-0	-0	-0	0	0.02
AMT_REQ_CREDIT_BUREAU_HOUR	0	0	-0	0	-0	-0	-0	0	0	-0	-0.01	-0	-0	-0	0	0	-0	-0	-0	0	-0	0	-0	1	0.23	0.01	0	0	0.01
AMT_REQ_CREDIT_BUREAU_DAY	0	0.01	0	0	0	-0	-0	0	0	0	0	-0	-0	-0	-0	-0	-0	0	-0.01	0	-0	0	-0	0.23	1	0.05	-0	-0.01	0.01
AMT_REQ_CREDIT_BUREAU_WEEK	0	0.01	-0	0.01	-0	-0	0	0.01	0	0.01	-0	0	-0	-0	-0	-0	-0	0	-0.02	-0	-0	-0	-0	0.01	0.05	1	-0.02	-0.02	0.05
AMT_REQ_CREDIT_BUREAU_MON	0	0.08	0.05	0.05	0.05	0.04	-0.01	0	0.01	0.02	0.02	0	0.01	0.01	-0.01	-0.01	-0.01	0.04	-0.02	0	0	0	-0	0	-0	-0.02	1	-0.01	0.09
AMT_REQ_CREDIT_BUREAU_QRT	-0	0.04	0.03	0.02	0.03	-0	0.02	0.02	-0	0.02	-0	-0.01	-0.01	-0.01	-0	-0.01	-0.01	0.01	-0.01	0.01	-0	0.01	0	0	-0.01	-0.02	-0.01	1	0.16
AMT_REQ_CREDIT_BUREAU_YEAR	-0.03	0.09	-0.01	0.01	-0.01	0	0.07	0.05	0.02	0.06	-0.02	-0.02	-0.02	-0.02	-0.02	-0.01	-0.01	-0	-0.06	0.03	0.02	0.03	0.02	0.01	0.01	0.05	0.09	0.16	1

We can infer that -:

- Credit amount is inversely proportional to the date of birth, which means credit amount is higher for low age and vice-versa,
- Credit amount is inversely proportional to the number of children client have, means Credit amount is higher for less children count client have and vice-versa,
- Income amount is inversely proportional to the number of children client have, means more income for less children client have and vice-versa,
- less children client have in densely populated area,
- Credit amount is higher, to densely populated area,
- The income is also higher in densely populated area.



Correlation for Defualters

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_REGISTRATION	YEARS_ID_PUBLISH	HOUR_APPR_PROCESS_START	REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	REG_CITY_NOT_LIVE_CITY	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	EXT_SOURCE_2	EXT_SOURCE_3	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	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	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1	-0	0	0.03	-0.01	-0.03	-0.27	-0.13	-0.13	0.03	-0.02	-0.03	-0.02	-0.01	0	0.05	0.05	-0.02	-0.02	0.03	-0	0.03	-0.01	0	-0.01	-0	-0.01	-0.01	-0.02
AMT_INCOME_TOTAL	-0	1	0.36	0.43	0.36	0.05	0.01	-0.04	-0.04	0.01	0.09	0.08	0.16	0.14	0.01	0.03	0.03	0.15	-0.07	0	-0.02	0	-0.02	0.01	0	0.03	0.09	0.04	0.12
AMT_CREDIT	0	0.36	1	0.81	0.98	0.04	0.16	0.08	0.02	0.05	0.03	0.01	0.03	0.03	-0.03	-0.03	-0.01	0.11	0.07	0.02	-0.02	0.02	-0.03	0	-0	0.01	0.05	-0.01	0.01
AMT_ANNUITY	0.03	0.43	0.81	1	0.8	0.02	0.03	-0	-0.04	0.01	0.03	0.03	0.06	0.06	-0	0.01	0.01	0.11	0.04	0.01	-0.02	0.01	-0.03	0.01	-0	0.03	0.06	0	0.01
AMT_GOODS_PRICE	-0.01	0.36	0.98	0.8	1	0.05	0.16	0.09	0.02	0.06	0.04	0.02	0.03	0.03	-0.03	-0.03	-0.01	0.12	0.08	0.02	-0.02	0.02	-0.03	0.01	-0	0.01	0.05	-0	0.01
REGION_POPULATION_RELATIVE	-0.03	0.05	0.04	0.02	0.05	1	0.05	0.02	0.03	0.02	0.11	-0.03	-0.01	0.01	-0.06	-0.04	-0.02	0.15	-0.01	0.03	0.05	0.03	0.04	-0	0	-0	0.05	-0.01	0.02
YEARS_BIRTH	-0.27	0.01	0.16	0.03	0.16	0.05	1	0.47	0.23	0.25	-0.07	-0.06	-0.06	-0.04	-0.17	-0.2	-0.11	0.12	0.17	-0.01	-0.01	-0.01	-0.01	-0.01	0	0.01	0	0.02	0.09
YEARS_EMPLOYED	-0.13	-0.04	0.08	-0	0.09	0.02	0.47	1	0.15	0.21	-0.04	-0.08	-0.12	-0.09	-0.14	-0.25	-0.17	0.07	0.12	0.01	-0	0.01	-0	-0.01	0	0.01	0	0.01	0.07
YEARS_REGISTRATION	-0.13	-0.04	0.02	-0.04	0.02	0.03	0.23	0.15	1	0.1	0.03	-0.02	-0.01	-0.01	-0.03	-0.07	-0.05	0.07	0.08	-0.02	-0.01	-0.02	-0.01	0	0	-0	-0.01	0.01	0.01
YEARS_ID_PUBLISH	0.03	0.01	0.05	0.01	0.06	0.02	0.25	0.21	0.1	1	-0.03	-0.03	-0.03	-0.02	-0.07	-0.07	-0.03	0.06	0.13	0.01	-0.01	0.01	-0.01	0	0.02	0	0.02	0.01	0.08
HOUR_APPR_PROCESS_START	-0.02	0.09	0.03	0.03	0.04	0.11	-0.07	-0.04	0.03	-0.03	1	0.05	0.07	0.05	0.01	0.02	0.01	0.14	-0.04	-0.02	-0	-0.02	-0	-0.01	-0.01	-0	0.03	0.01	-0.01
REG_REGION_NOT_LIVE_REGION	-0.03	0.08	0.01	0.03	0.02	-0.03	-0.06	-0.08	-0.02	-0.03	0.05	1	0.5	0.07	0.32	0.15	-0.01	0.01	-0.05	-0.03	-0	-0.03	-0	-0.01	-0.01	-0.01	0.02	-0	-0.03
REG_REGION_NOT_WORK_REGION	-0.02	0.16	0.03	0.06	0.03	-0.01	-0.06	-0.12	-0.01	-0.03	0.07	0.5	1	0.85	0.14	0.24	0.19	0.02	-0.03	-0.04	-0.02	-0.04	-0.02	-0	-0	-0.01	0.01	-0.01	-0.02
LIVE_REGION_NOT_WORK_REGION	-0.01	0.14	0.03	0.06	0.03	0.01	-0.04	-0.09	-0.01	-0.02	0.05	0.07	0.85	1	-0	0.18	0.24	0.02	-0.01	-0.03	-0.02	-0.03	-0.02	0	-0	-0	0.01	-0	-0.01
REG_CITY_NOT_LIVE_CITY	0	0.01	-0.03	-0	-0.03	-0.06	-0.17	-0.14	-0.03	-0.07	0.01	0.32	0.14	-0	1	0.47	-0.01	-0.05	-0.05	-0.02	0.01	-0.02	0.02	-0	-0.01	-0.01	-0	0	-0.03
REG_CITY_NOT_WORK_CITY	0.05	0.03	-0.03	0.01	-0.03	-0.04	-0.2	-0.25	-0.07	-0.07	0.02	0.15	0.24	0.18	0.47	1	0.78	-0.07	-0.05	-0.02	-0	-0.02	-0	-0.01	-0.01	-0.01	-0	-0.01	-0.02
LIVE_CITY_NOT_WORK_CITY	0.05	0.03	-0.01	0.01	-0.01	-0.02	-0.11	-0.17	-0.05	-0.03	0.01	-0.01	0.19	0.24	-0.01	0.78	1	-0.05	-0.02	-0.01	-0.01	-0.01	-0.02	-0.01	-0	-0.01	-0	-0.02	-0.01
EXT_SOURCE_2	-0.02	0.15	0.11	0.11	0.12	0.15	0.12	0.07	0.07	0.06	0.14	0.01	0.02	0.02	-0.05	-0.07	-0.05	1	0.08	0.02	-0.01	0.02	-0.01	-0	-0	0.01	0.05	0.01	0.04
EXT_SOURCE_3	-0.02	-0.07	0.07	0.04	0.08	-0.01	0.17	0.12	0.08	0.13	-0.04	-0.05	-0.03	-0.01	-0.05	-0.05	-0.02	0.08	1	-0.01	-0.03	-0.01	-0.02	0	0	-0.02	-0.03	-0.04	-0.04
OBS_30_CNT_SOCIAL_CIRCLE	0.03	0	0.02	0.01	0.02	0.03	-0.01	0.01	-0.02	0.01	-0.02	-0.03	-0.04	-0.03	-0.02	-0.02	-0.01	0.02	-0.01	1	0.42	1	0.35	-0	-0	0	0.01	0	0.04
DEF_30_CNT_SOCIAL_CIRCLE	-0	-0.02	-0.02	-0.02	-0.02	0.05	-0.01	-0	-0.01	-0.01	-0	-0	-0.02	-0.02	0.01	-0	-0.01	-0.01	-0.03	0.42	1	0.43	0.86	0	0	-0.01	-0	-0	-0
OBS_60_CNT_SOCIAL_CIRCLE	0.03	0	0.02	0.01	0.02	0.03	-0.01	0.01	-0.02	0.01	-0.02	-0.03	-0.04	-0.03	-0.02	-0.02	-0.01	0.02	-0.01	1	0.43	1	0.35	-0	-0	0	0.01	0	0.04
DEF_60_CNT_SOCIAL_CIRCLE	-0.01	-0.02	-0.03	-0.03	-0.03	0.04	-0.01	-0	-0.01	-0.01	-0	-0	-0.02	-0.02	0.02	-0	-0.02	-0.01	-0.02	0.35	0.86	0.35	1	-0	0	-0	-0.01	-0	-0.01
AMT_REQ_CREDIT_BUREAU_HOUR	0	0.01	0	0.01	0.01	-0	-0.01	-0.01	0	0	-0.01	-0.01	-0	0	-0	-0.01	-0.01	-0	0	-0	0	-0	-0	1	0.25	0.01	-0	0.01	0.02
AMT_REQ_CREDIT_BUREAU_DAY	-0.01	0	-0	-0	-0	0	0	0	0	0.02	-0.01	-0.01	-0	-0	-0.01	-0.01	-0	-0	0	-0	0	-0	0	0.25	1	0.04	-0.01	0	0.03
AMT_REQ_CREDIT_BUREAU_WEEK	-0	0.03	0.01	0.03	0.01	-0	0.01	0.01	-0	0	-0	-0.01	-0.01	-0	-0.01	-0.01	-0.01	0.01	-0.02	0	-0.01	0	-0	0.01	0.04	1	-0.01	-0	0.06
AMT_REQ_CREDIT_BUREAU_MON	-0.01	0.09	0.05	0.06	0.05	0.05	0	0	-0.01	0.02	0.03	0.02	0.01	0.01	-0	-0	-0	0.05	-0.03	0.01	-0	0.01	-0.01	-0	-0.01	-0.01	1	0.02	0.11
AMT_REQ_CREDIT_BUREAU_QRT	-0.01	0.04	-0.01	0	-0	-0.01	0.02	0.01	0.01	0.01	0.01	-0	-0.01	-0	0	-0.01	-0.02	0.01	-0.04	0	-0	0	-0	0.01	0	-0	0.02	1	0.18
AMT_REQ_CREDIT_BUREAU_YEAR	-0.02	0.12	0.01	0.01	0.01	0.02	0.09	0.07	0.01	0.08	-0.01	-0.03	-0.02	-0.01	-0.03	-0.02	-0.01	0.04	-0.04	0.04	-0	0.04	-0.01	0.02	0.03	0.06	0.11	0.18	1

We can infer that -:

- The client's permanent address does not match contact address are having less children and vice-versa.
- The client's permanent address does not match work address are having less children and vice-versa.

From both the Heatmaps we can conclude that

The top 5 correlation for Defaulters are as follows,

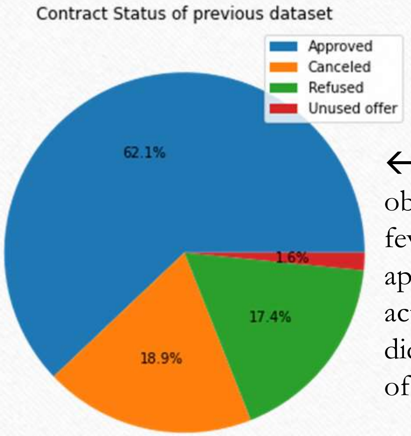
- Between Amount credit and amount of goods price,
- Between Amount credit and amount of annuity,
- Between Amount annuity and amount of goods price
- Between (Working address not being working address) and (Living region not being working region)
- Between (Working city not being working city) and (Living city not being working city)

The top 5 correlation for Non-Defaulters are as follows

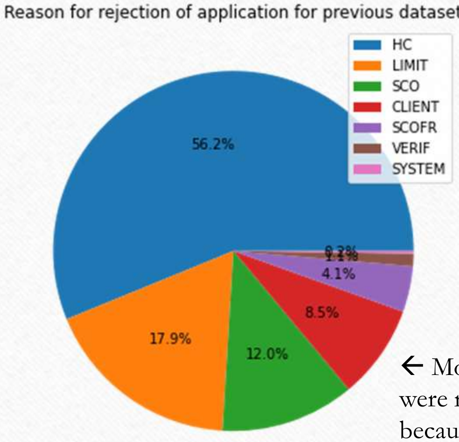
- Between Amount credit and amount of goods price,
- Between Amount credit and amount of annuity,
- Between Amount annuity and amount of goods price
- Between number of observation of client's social surroundings defaulted on 30 DPD and 60 DPD
- Between (Working address not being working address) and (Living region not being working region)

PREVIOUS DATA

Once we were done with analyzing and drawing insight from Current data, it was time to move the previous data into the picture, so we did the same process of cleaning, and checked for outliers, once we were done with it we started analyzing the dataset, and these were the insights that we could gather -:

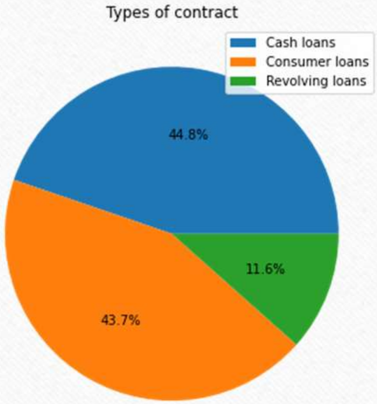


← As we can observe very few of the applicants actually didn't use the offer



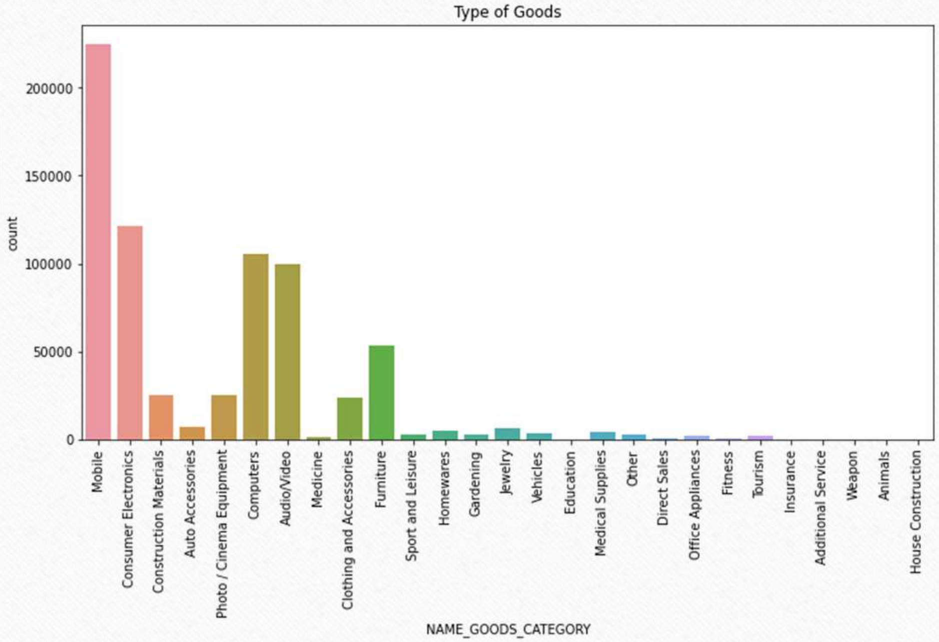
← Most people were rejected because of 'HC'

AS we can see → majority of loans are of either Cash type or consumer type, with very few as Revolving type



UNIVARIATE ANALYSYS ON PREVIOUS DATA

-- Bar chart--

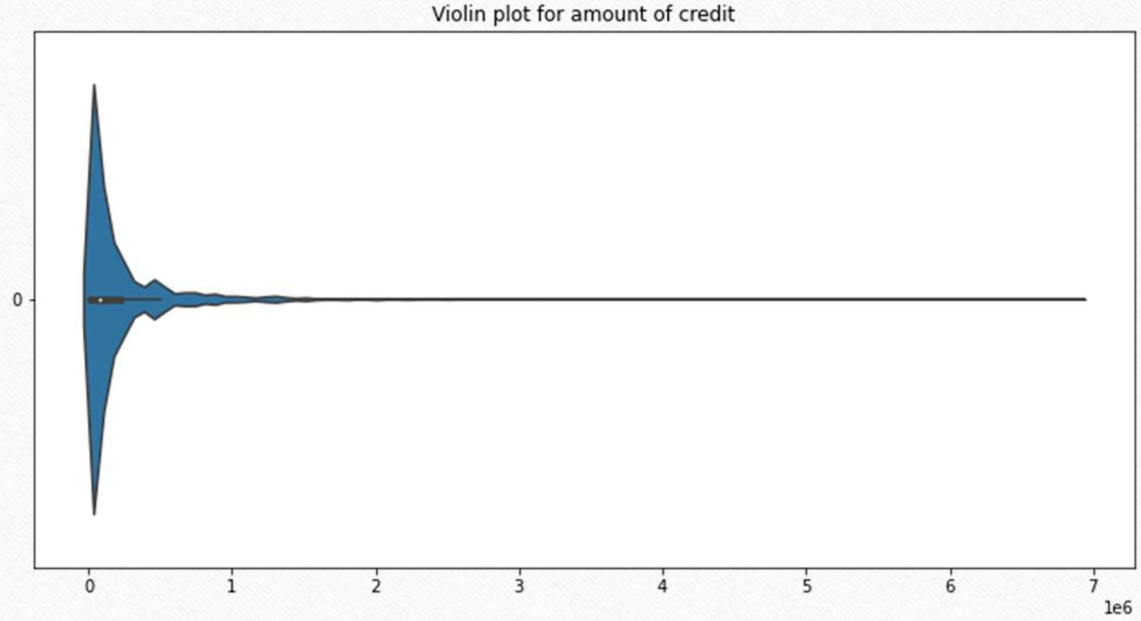


Here we plotted a bar chart for count of good category for which a person applies for a loan

And we inferred that see majority of goods for which the client applied the loan for were mobiles, consumer electronics, computers and audio/video.

UNIVARIATE ANALYSYS ON PREVIOUS DATA

-- Violin Plot--



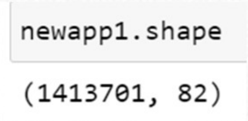
Using a violin plot we can check for outliers as evident in the figure,

We can see that though there were certain outliers present in 'Amount of credit' for which the person has applied a loan for,

These outliers are nothing out of the ordinary and can simply be tackled by binning/capping the column values.

MERGING BOTH THE DATASETS

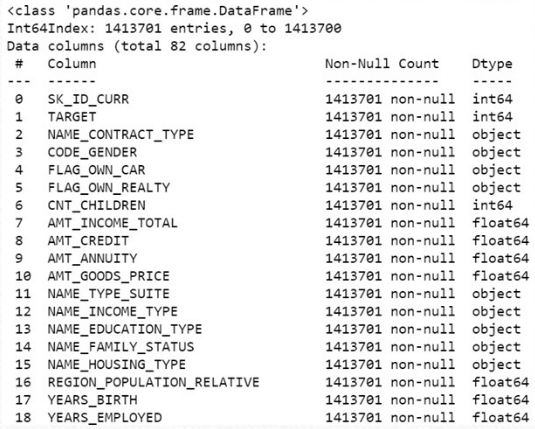
Now that we have dealt with outliers and have done all required cleaning, its time to merge both the datasets and draw our inferences as per the problem statement, but before beginning to plot the merged data set, we need to look at both attributes and statical summary of our merged data frame,



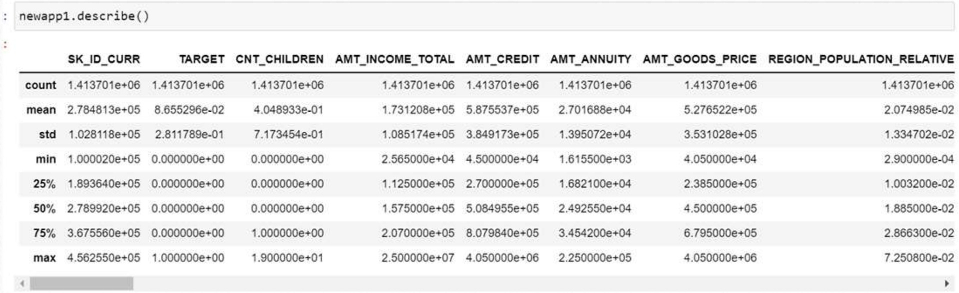
```
newapp1.shape
```

```
(1413701, 82)
```

As we can see there are 1413701 rows and 82 columns



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1413701 entries, 0 to 1413700
Data columns (total 82 columns):
 #   Column                      Non-Null Count    Dtype
---  ------                      --------------    -----
 0   SK_ID_CURR                  1413701 non-null  int64
 1   TARGET                      1413701 non-null  int64
 2   NAME_CONTRACT_TYPE          1413701 non-null  object
 3   CODE_GENDER                 1413701 non-null  object
 4   FLAG_OWN_CAR                1413701 non-null  object
 5   FLAG_OWN_REALTY             1413701 non-null  object
 6   CNT_CHILDREN                1413701 non-null  int64
 7   AMT_INCOME_TOTAL            1413701 non-null  float64
 8   AMT_CREDIT                  1413701 non-null  float64
 9   AMT_ANNUITY                 1413701 non-null  float64
 10  AMT_GOODS_PRICE             1413701 non-null  float64
 11  NAME_TYPE_SUITE             1413701 non-null  object
 12  NAME_INCOME_TYPE            1413701 non-null  object
 13  NAME_EDUCATION_TYPE         1413701 non-null  object
 14  NAME_FAMILY_STATUS          1413701 non-null  object
 15  NAME_HOUSING_TYPE           1413701 non-null  object
 16  REGION_POPULATION_RELATIVE  1413701 non-null  float64
 17  YEARS_BIRTH                 1413701 non-null  float64
 18  YEARS_EMPLOYED              1413701 non-null  float64
```



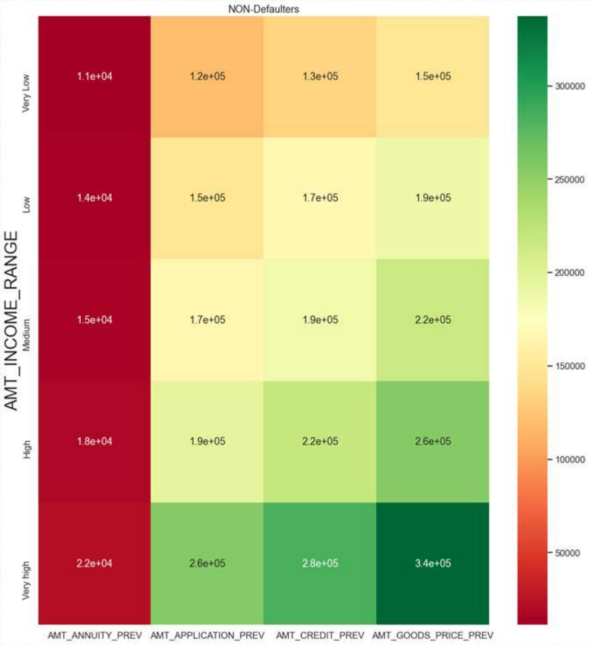
```
newapp1.describe()
```

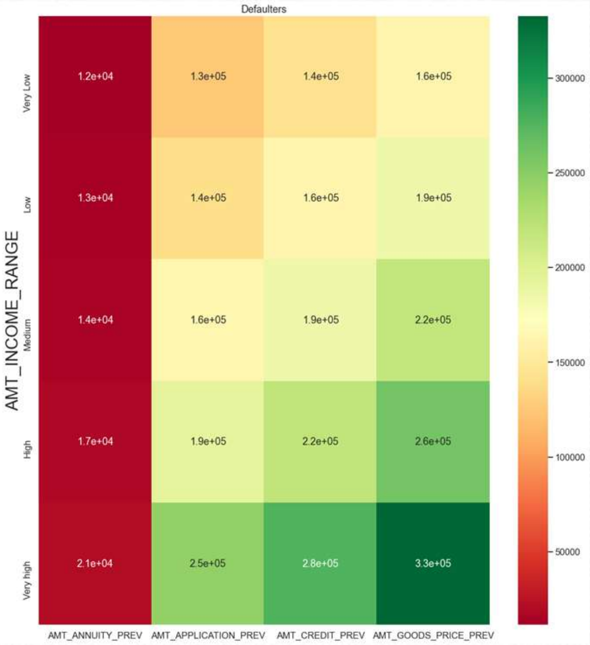
	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE
count	1.413701e+06	1.413701e+06	1.413701e+06	1.413701e+06	1.413701e+06	1.413701e+06	1.413701e+06	1.413701e+06
mean	2.784813e+05	8.655296e-02	4.048933e-01	1.731208e+05	5.875537e+05	2.701688e+04	5.276522e+05	2.074985e-02
std	1.028118e+05	2.811789e-01	7.173454e-01	1.085174e+05	3.849173e+05	1.395072e+04	3.531028e+05	1.334702e-02
min	1.000020e+05	0.000000e+00	0.000000e+00	2.565000e+04	4.500000e+04	1.615500e+03	4.050000e+04	2.900000e-04
25%	1.893640e+05	0.000000e+00	0.000000e+00	1.125000e+05	2.700000e+05	1.682100e+04	2.385000e+05	1.003200e-02
50%	2.789920e+05	0.000000e+00	0.000000e+00	1.575000e+05	5.084955e+05	2.492550e+04	4.500000e+05	1.885000e-02
75%	3.675560e+05	0.000000e+00	1.000000e+00	2.070000e+05	8.079840e+05	3.454200e+04	6.795000e+05	2.866300e-02
max	4.562550e+05	1.000000e+00	1.900000e+01	2.500000e+07	4.050000e+06	2.250000e+05	4.050000e+06	7.250800e-02

All columns seems to be associated with correct datatype,
And numerical summary seems to be impeccable.

BIVARIATE ANALYSIS ON MERGED DATASET

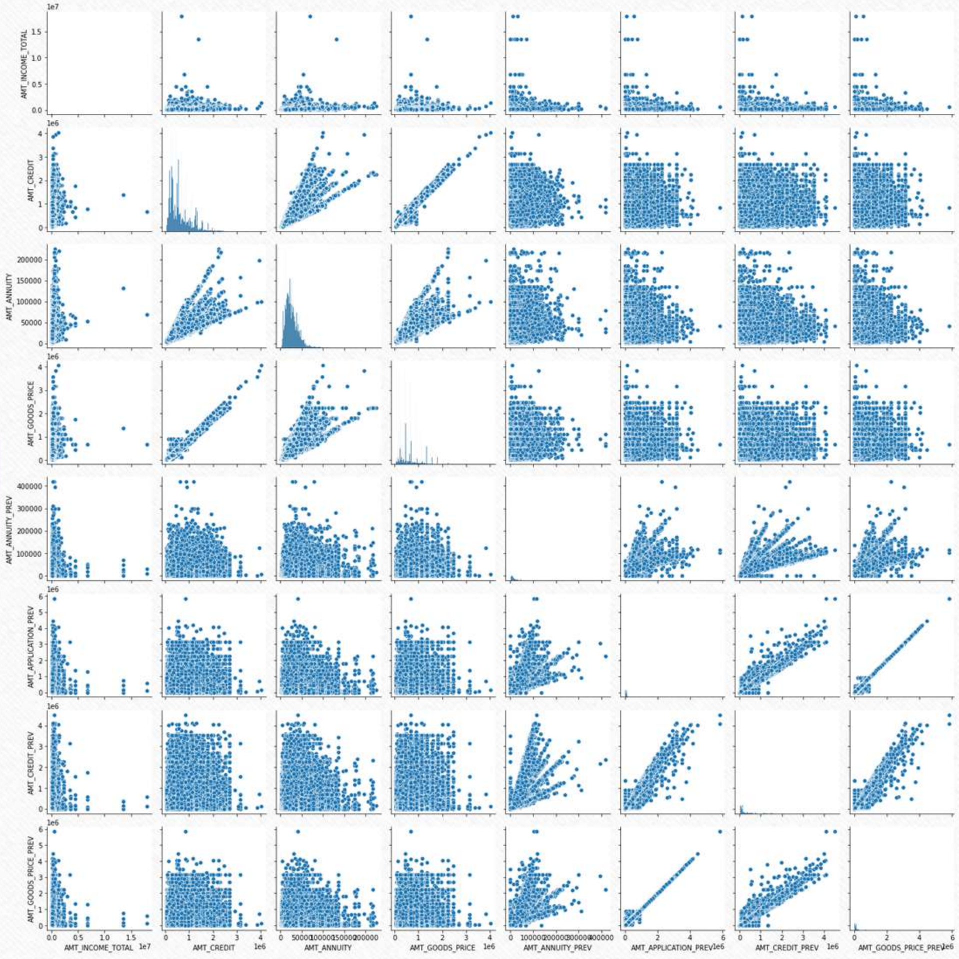
--HEAT MAPS--





When we plotted Heat Maps between Amount of Income in various range,
With different sorts of attributes related to Amount, segmented by Non-defaulters and defaulters,

We concluded that for defaulters all attributes related to amounts were higher than that of Non-defaulters in case of low income group, while the opposite is true for clients in very high income group.

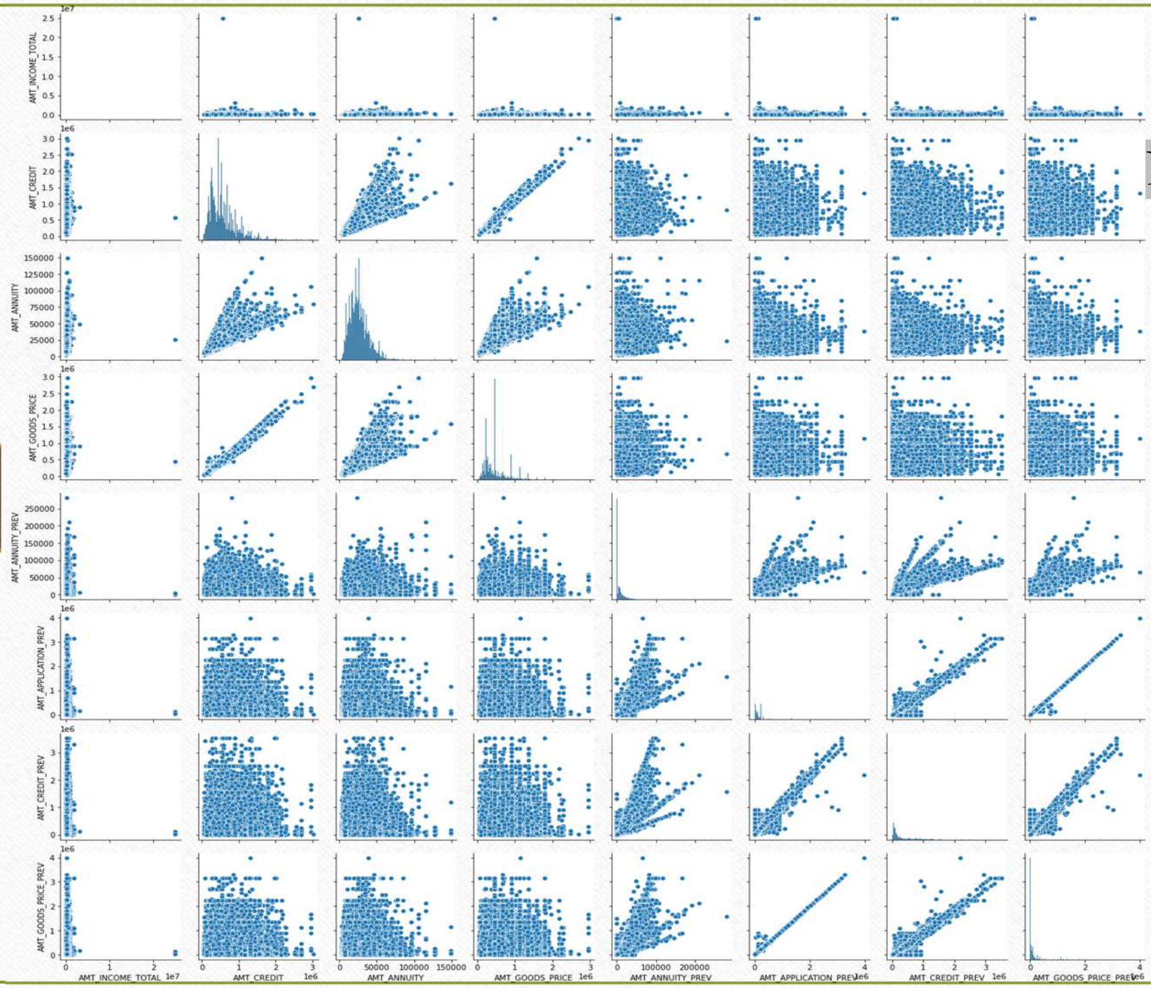


BIVARIATE ANALYSIS ON MERGED DATASET USING PAIRPLOTS FOR NON-DEFAULTERS

We can see from the figure that,

Top 5 correlations are - :

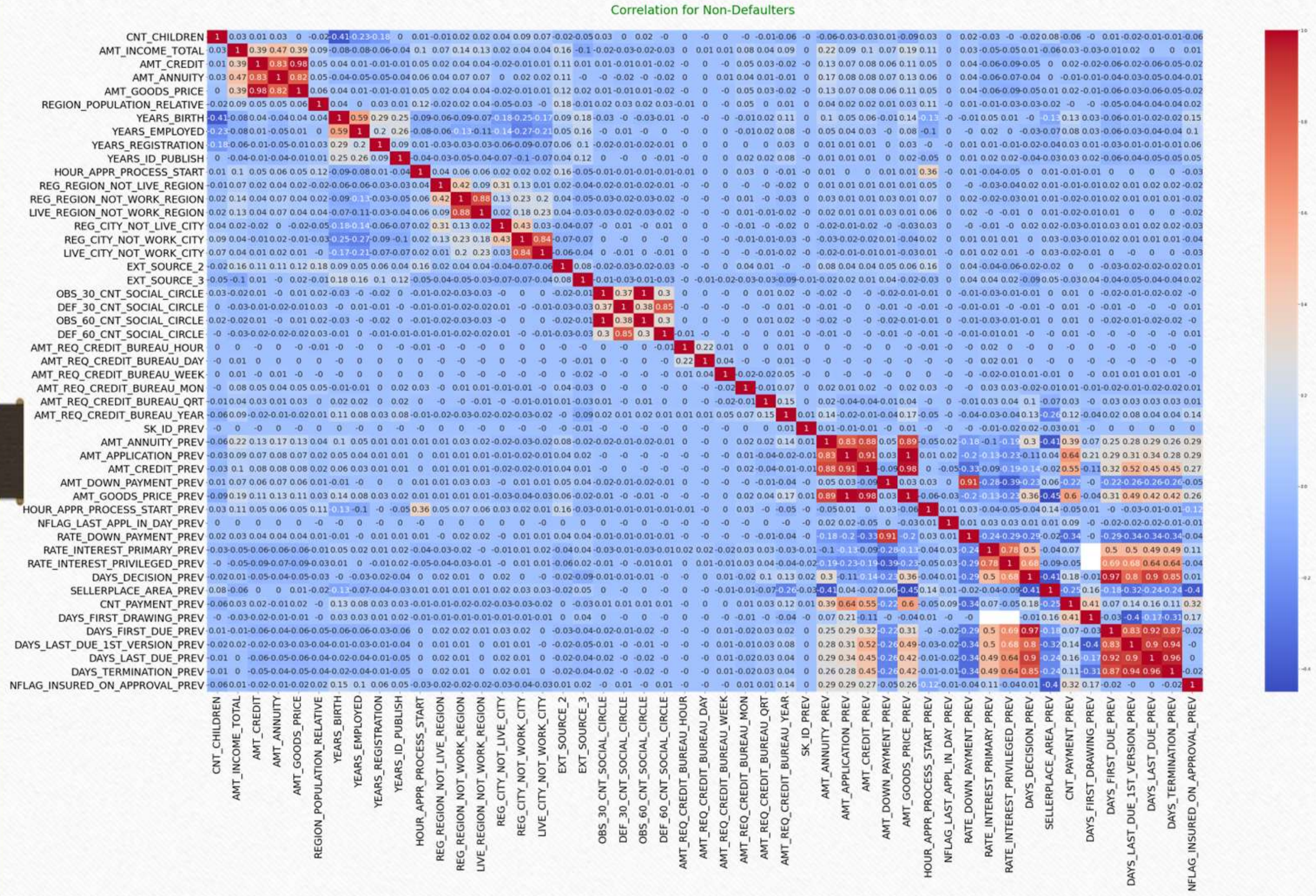
- Amount of credit and Amount of goods price,
- Amount of credit and Amount of Annuity,
- Amount of Annuity and Amount of goods price
- Amount of Credit asked by client previously and Currently
- Amount of Credit asked by client previously and Amount of price of goods previously



BIVARIATE ANALYSIS ON MERGED DATASET USING PAIRPLOTS FOR DEFAULTERS

We can see from the figure that,

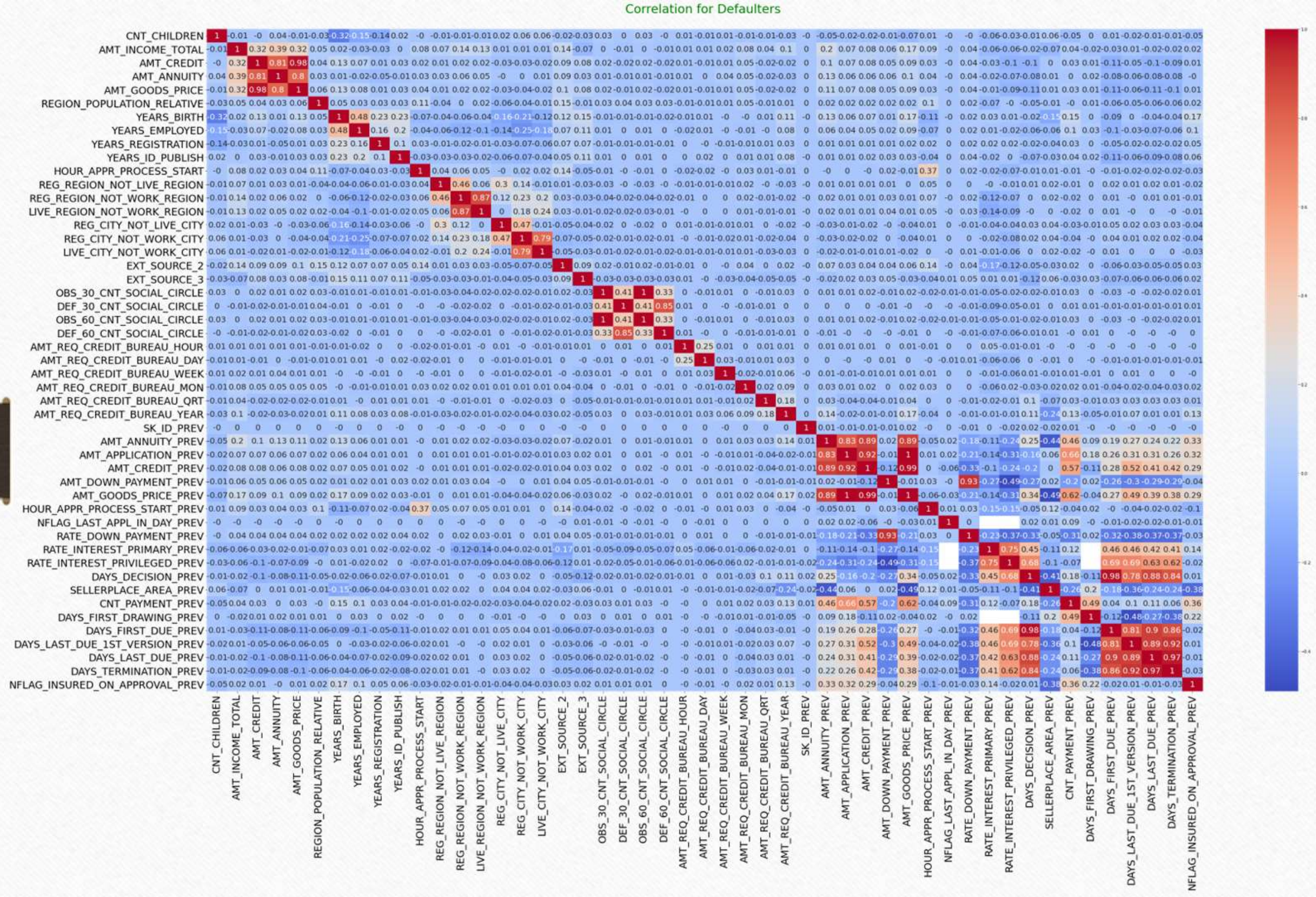
Top 5 correlations are same as that of non-defaulters



HEAT MAPS For Non-Defaulters, for all Numerical columns

From the figure it is evident that top 5 correlations are -:

- Amount of credit and Amount of goods price,
- Amount of credit and Amount of Annuity,
- Amount of Annuity and Amount of goods price
- Amount of Credit asked by client previously and Currently
- Amount of Credit asked by client previously and Amount of price of goods previously



HEAT MAPS For Defaulters, for all Numerical columns

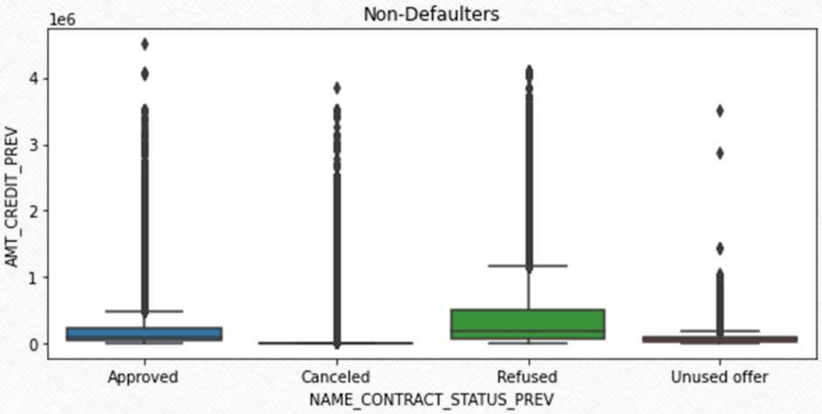
From the figure it is evident that top 5 correlations are -:

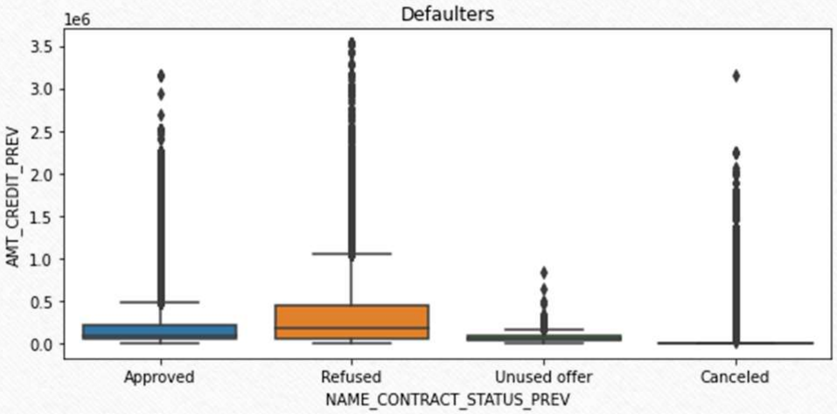
- Amount of credit and Amount of goods price,
- Amount of credit and Amount of Annuity,
- Amount of Annuity and Amount of goods price
- Amount of Credit asked by client previously and Currently
- Amount of Credit asked by client previously and Amount of price of goods previously

MULTIVARIATE ANALYSIS OF MERGED DATASET

--BOX PLOT--

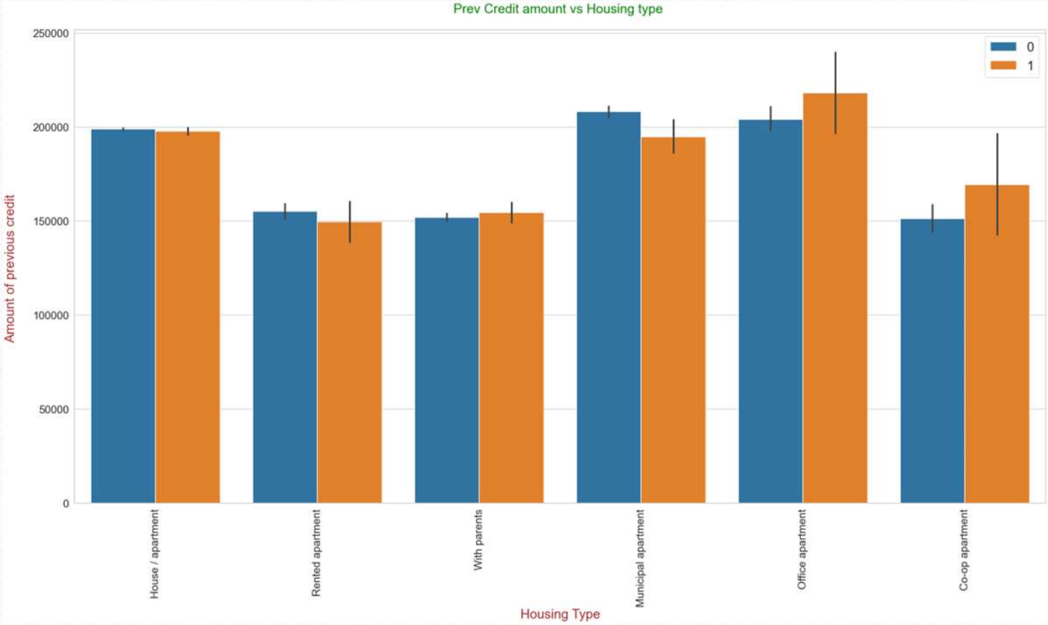
While performing a box plot on both defaulters and non-defaulters, on basis of the amount of credit requested by them, and the outcome it entailed we reached the conclusion that both defaulters and non-defaulters who were Refused the most had requested higher credits than their counterparts.





MULTIVARIATE ANALYSIS OF MERGED DATASET

--Bar Chart--



Here when we plotted a bar chart for Amount of previous credit and housing type on basis of them being defaulter or not,

We learned that or Housing type, office apartment is having higher credit of Non-Defaulters and co-op apartment is having higher credit of Defaulters. So, we can conclude that bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment. Bank can focus mostly on housing type with parents or House/apartment or municipal apartments for successful payments.

CONCLUSION

FOCUS

- Banks should focus more on contract type 'Student', 'Pensioner' and 'Businessman' with housing 'type other than 'co-op apartment' for successful payments.
- Banks should focus less on income type 'working' as they are having most number of unsuccessful payments.
- Also with loan purpose 'Repair' is having higher number of unsuccessful payments on time.
- Get as much as clients from housing type 'With parents' as they are having least number of unsuccessful payments.

MOST LIKELY TO DEFAULT

- Mostly Single people,
- Mostly adults and young adults,
- Clients pursuing revolving loan are more likely to default than client perusing other types of loans,
- Clients with Secondary Education are more likely to be defaulters
- Clients having low income &
- Clients who were rejected for the loan before.

LEAST LIKELY TO DEFAULT

- Mostly married people,
- Mostly middle aged and senior citizens,
- Clients pursuing other types of loans rather than revolving loan,
- People with higher education,
- Clients having high income &
- Clients who were approved for the loan before.