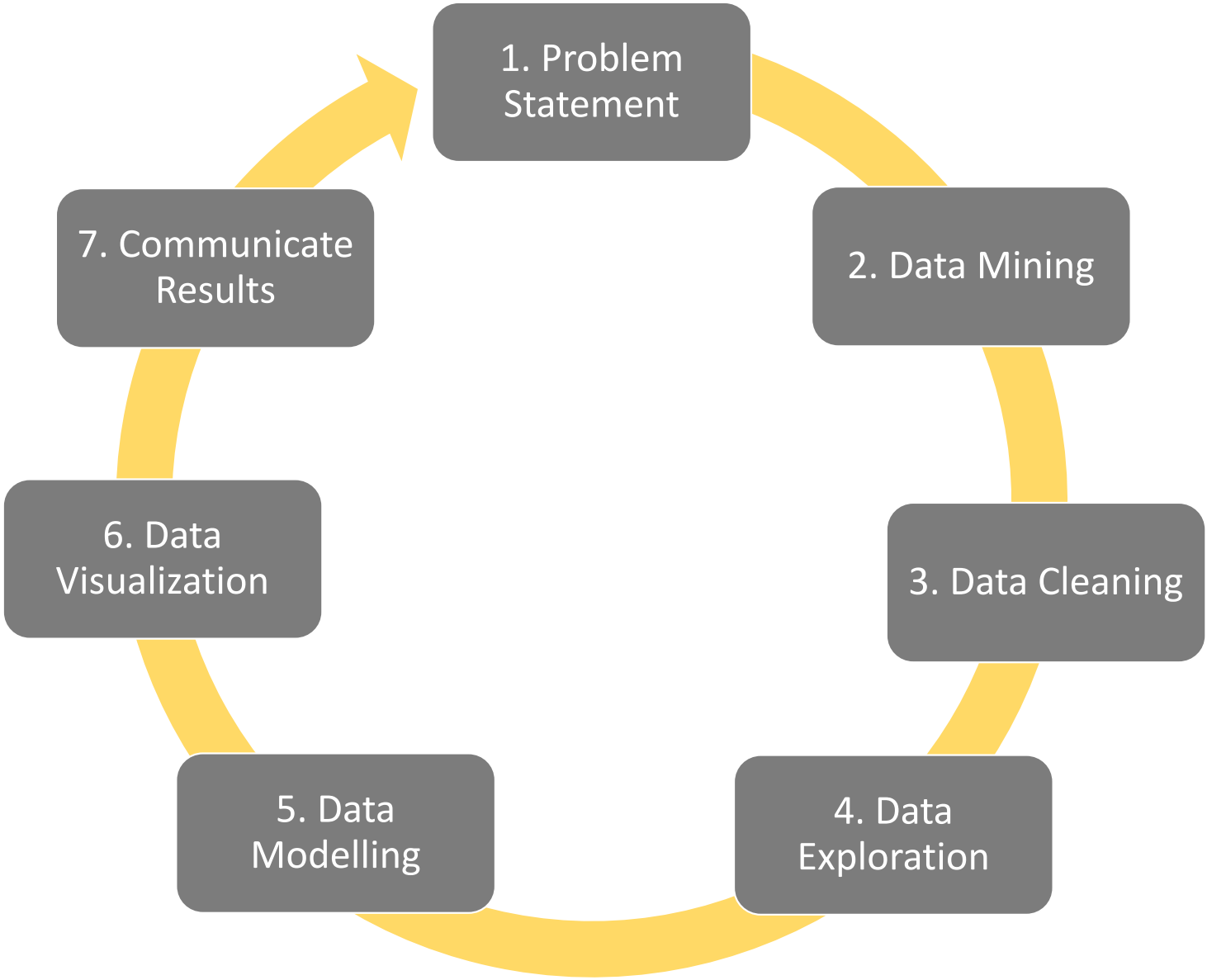
Credit EDA case study

Presented by:

- Mohammad Shahid Rashid (DCS27)
- Kirk Xyrrus Villanueva (DCS27)

EDA Life cycle

1. Problem Statement:-
   Business understanding and define objectives for the problem that needs to be handled
2. Data Mining:-
   Gather and scrap the data which is needed for this case study.
3. Data Cleaning:-
   Fix the inconsistencies within the data and handle missing values
4. Data Exploration
   Form hypotheses about your defined problem by visually analysing the data
5. Data Modelling
   Extract important features and construct a more meaningful one with the raw data.
6. Data Visualization:-
   Prepare the finding with plots.
7. Communicate Results
   Based on such far progress, summarized the results with appropriate reason. Few Recommendation



1. Problem Statement

- If applicants are **likely to repay** the loan.

Approve
- Bank loan business will be stable

Reject
- Loss of business to the company (Interest loss)

- If applicants are **not likely to repay** the loan.

Approve
- Financial loss for the company. (Credit loss)

Reject
- Good decision to save loan entity business.

Report the variable that can help the bank to identify if the applicant could be default or not

2. Data Mining

Import Library

```
# Importing the libraries

import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

Read csv file

```
# Importing the files to be used for analysis

ca = pd.read_csv('application_data.csv')
pa = pd.read_csv('previous_application.csv')
```

Data structure check

- ca.head()
- ca.shape
- ca.describe()
- ca.info()
- ca.duplicated()

Null Value check

- ca.isnull().sum()*100

Outliers

- Max and Min value
- Box plot check

3. Data Cleaning

Drop high Null values → Impute smaller value → Identify outlier → Datatypes & negative values → Binnng Continuous Variable

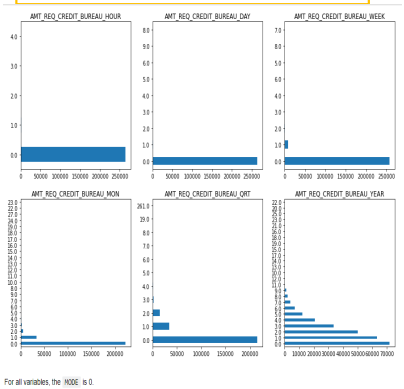
Variable >50% null values dropped from application_data

OWN_CAR_AGE
EXT_SOURCE_1
APARTMENTS_AVG
BASEMENTAREA_AVG
YEARS_BUILD_AVG
COMMONAREA_AVG
ELEVATORS_AVG
ENTRANCES_AVG
FLOORSMIN_AVG
LANDAREA_AVG
LIVINGAPARTMENTS_AVG
LIVINGAREA_AVG
NONLIVINGAPARTMENTS_AVG
NONLIVINGAREA_AVG
APARTMENTS_MODE
BASEMENTAREA_MODE
YEARS_BUILD_MODE
COMMONAREA_MODE
ELEVATORS_MODE
ENTRANCES_MODE
FLOORSMIN_MODE
LANDAREA_MODE
LIVINGAPARTMENTS_MODE
LIVINGAREA_MODE
NONLIVINGAPARTMENTS_MODE
NONLIVINGAREA_MODE
APARTMENTS_MEDI
BASEMENTAREA_MEDI
YEARS_BUILD_MEDI
COMMONAREA_MEDI
ELEVATORS_MEDI
ENTRANCES_MEDI
FLOORSMIN_MEDI
LANDAREA_MEDI
LIVINGAPARTMENTS_MEDI
LIVINGAREA_MEDI
NONLIVINGAPARTMENTS_MEDI
NONLIVINGAREA_MEDI
FONDKAPREMONT_MODE
HOUSETYPE_MODE
WALLSMATERIAL_MODE

MEDIAN
- AMT_ANNUITY
- AMT_GOODS_PRICE

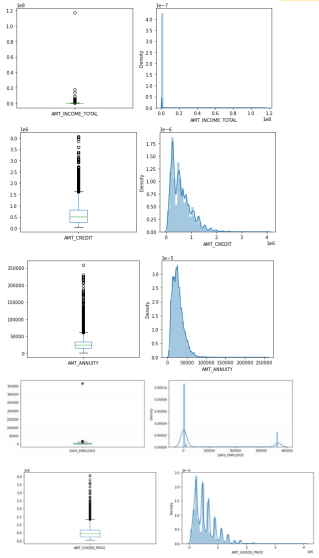
MODE
- NAME_TYPE_SUITE
- CNT_FAM_MEMBERS
- OBS_30_CNT_SOCIAL_CIRCLE
- DEF_30_CNT_SOCIAL_CIRCLE

MEAN
- EXT_SOURCE_2



For all variables, the MODE is 0.

Handling Outliers
- AMT_INCOME_TOTAL (90th %ile)
- AMT_CREDIT (95th %ile)
- AMT_ANNUITY (97th %ile)
- AMT_GOODS_PRICE (95th %ile)
- DAYS_EMPLOYED (70th %ile)



Convert into absolute value(*positive*)
- DAYS_BIRTH
- DAYS_EMPLOYED
- DAYS_REGISTRATION
- DAYS_ID_PUBLISH
- DAYS_LAST_PHONE_CHANGE

	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	DAYS_LAST_PHONE_CHANGE
0	-9461	-637	-3648.00	-2120	-1134.00
1	-16765	-1188	-1186.00	-291	-828.00
2	-19046	-225	-4260.00	-2531	-815.00
3	-19005	-3039	-9833.00	-2437	-617.00
4	-19932	-3038	-4311.00	-3458	-1106.00
5	-16941	-1588	-4970.00	-477	-2536.00
6	-13778	-3130	-1213.00	-619	-1562.00
7	-18850	-449	-4597.00	-2379	-1070.00
8	-20099	365243	-7427.00	-3514	0.00
9	-14469	-2019	-14437.00	-3992	-1673.00

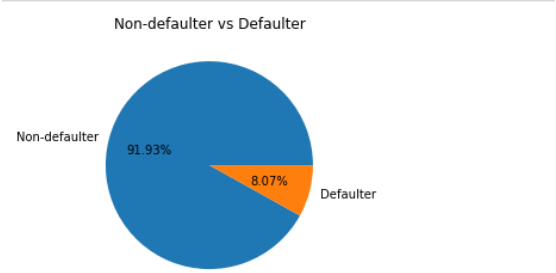
Binning based on quantiles

INCOME_CATEGORY *(AMT_INCOME_TOTAL)*
- Low Income<= 112500
- 112500 <Average Income<= 147150
- 147150 <Moderate Income<= 202500
- High Income>202500

CREDIT_CATEGORY *(AMT_CREDIT)*
- Low-End Credit <= 270000
- 270000 <Moderate Credit <= 513531
- 513531 <Moderately-High Credit <= 808650
- High-End Credit >808650

4. Data Exploration

Checking **Imbalance** Percentage



There is a huge imbalance between the **Non-defaulters** and **Defaulters**.

Dividing the Dataset into Non-Defaulters and Defaulters Dataframe

```
ca_ndef = ca[ca['TARGET'] == 0]
ca_def = ca[ca['TARGET'] == 1]
```

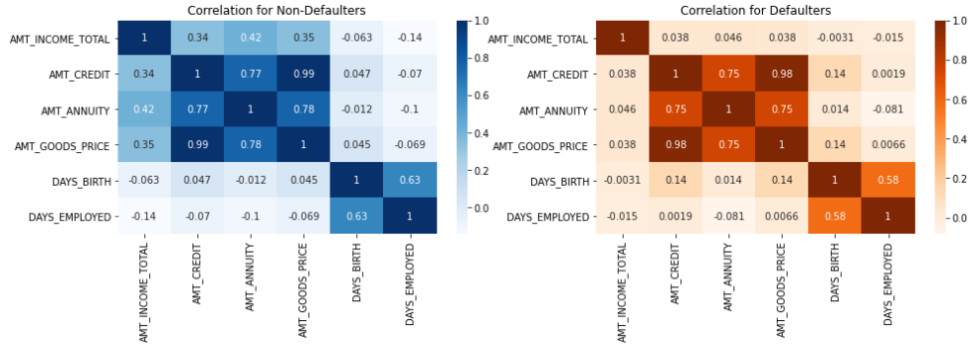
```
ca_ndef.head()
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	A
1	100003	0	Cash loans	F	N	N	0	270000.00	
2	100004	0	Revolving loans	M	Y	Y	0	67500.00	
3	100006	0	Cash loans	F	N	Y	0	135000.00	
4	100007	0	Cash loans	M	N	Y	0	121500.00	
5	100008	0	Cash loans	M	N	Y	0	99000.00	

```
ca_def.head()
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	A
0	100002	1	Cash loans	M	N	Y	0	202500.00	
26	100031	1	Cash loans	F	N	Y	0	112500.00	
40	100047	1	Cash loans	M	N	Y	0	202500.00	
42	100049	1	Cash loans	F	N	N	0	135000.00	
81	100096	1	Cash loans	F	N	Y	0	81000.00	

Correlation for **Numerical** Variables



Based on the heatmap, the highest correlation between the two datasets are on the same variables.

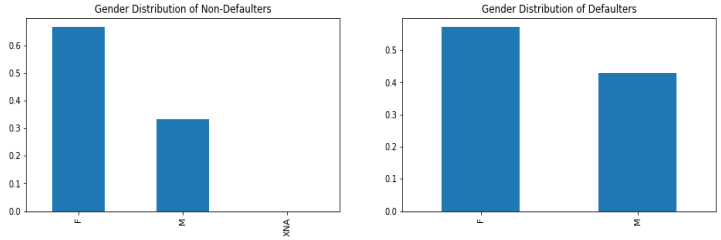
Based on the heatmap, the highest correlation in the two datasets are of the same variables.

AMT_CREDIT vs AMT_ANNUITY
AMT_ANNUITY vs AMT_GOODS_PRICE
AMT_CREDIT vs AMT_GOODS_PRICE
DAYS_BIRTH vs DAYS_EMPLOYED

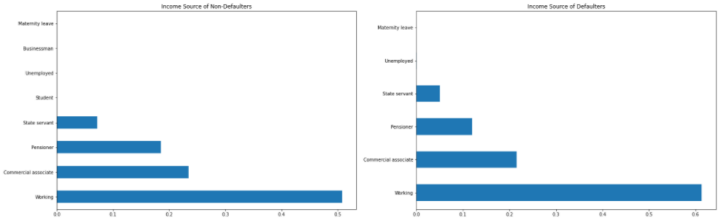
5. Data Modeling

Univariate Analysis for Categorical Variables

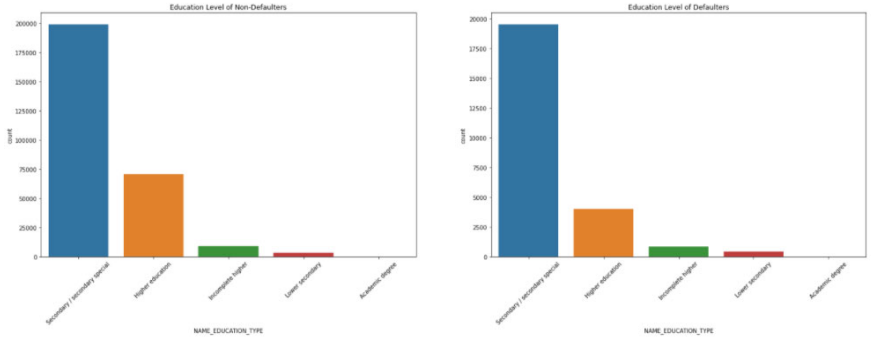
CODE_GENDER :- More females tend to apply for loans irrespective of being Non-Defaulter or Defaulter.



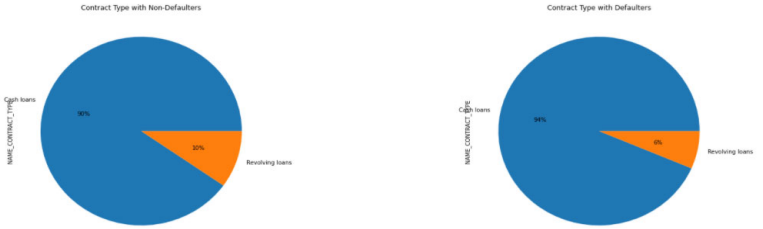
NAME_INCOME_TYPE :- Working individuals represent the highest part of the distribution for those applying for a loan.



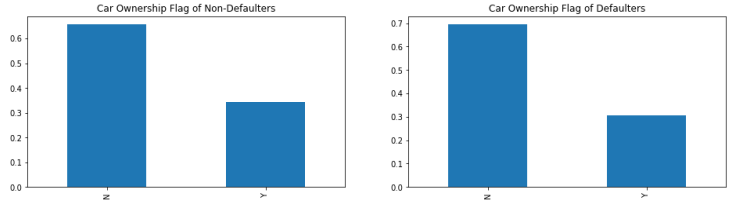
NAME_EDUCATION_TYPE :- The majority of those applying for loans are with a secondary level of educational attainment.



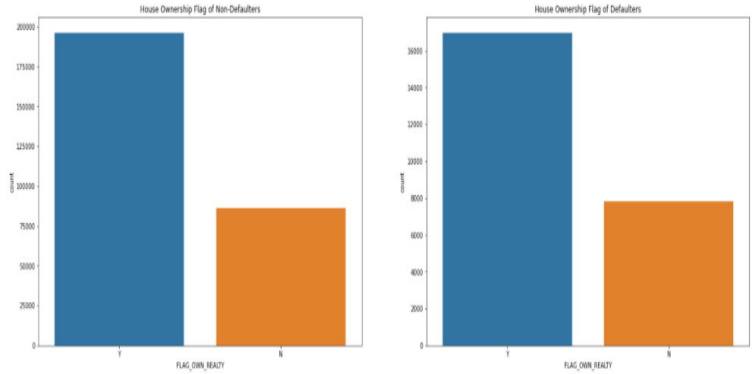
NAME_CONTRACT_TYPE :- Cash loans are more often applied to as compared to Revolving loans.



FLAG_OWN_CAR :- The majority of the new loan applicants do not own a car.



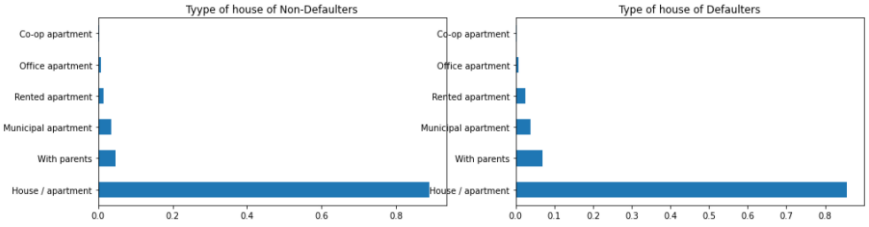
FLAG_OWN_REALTY :- On the contrary, with car ownership, the majority of the loan applicants have their residence.



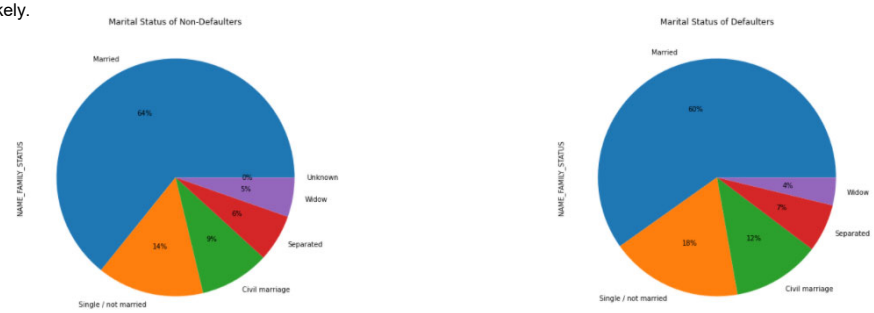
Continuation...

Univariate Analysis for Categorical Variables

NAME_HOUSING_TYPE:- Most of those applying for loans live in an apartment

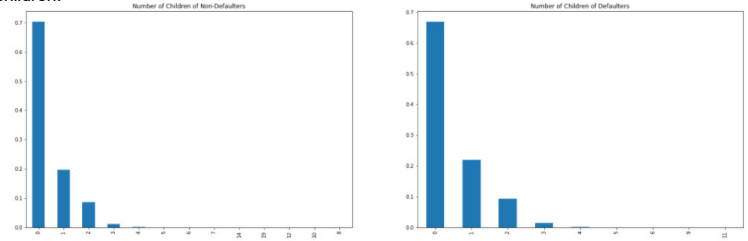


NAME_FAMILY_STATUS:- Widowers and separated individuals tend to apply for a loan less likely.



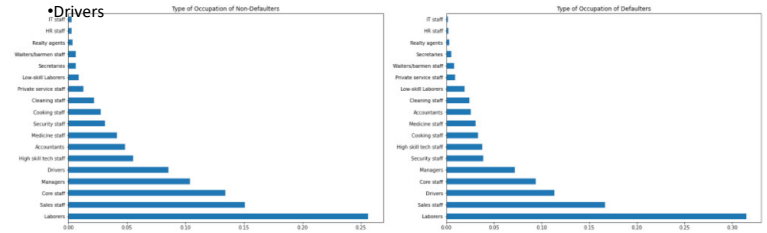
In general, there are no apparent differences (visually) in the characteristics of those who have defaulted their loans with that of a non-defaulter. The patterns found in the variables are pretty similar.

CNT_CHILDREN :- Applicants with a less number of children applies more frequently than those with a larger family size. Highest frequency of loan application comes from those without any children.

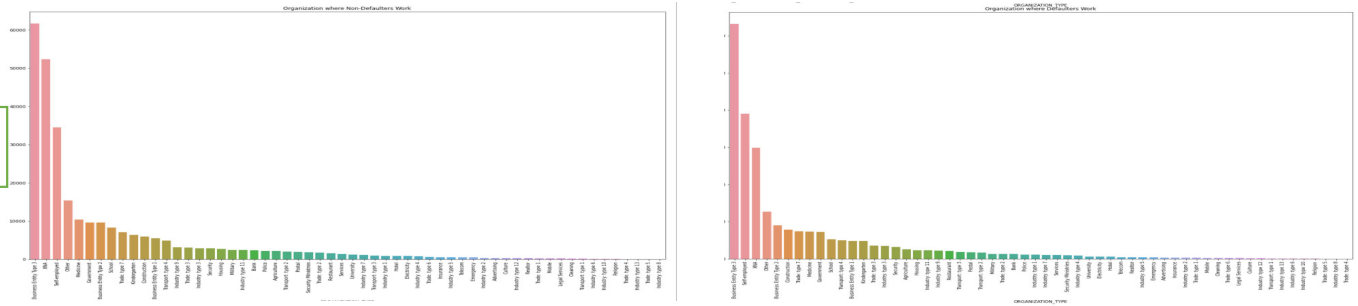


OCCUPATION_TYPE:- The top five occupations of those taking loans for both Non-Defaulters and Defaulters are as follows:

- Laborers
- Sales Staff
- Core Staff
- Managers
- Drivers



ORGANIZATION_TYPE:- The highest frequency of loan applications comes from the Business Entity organization type.

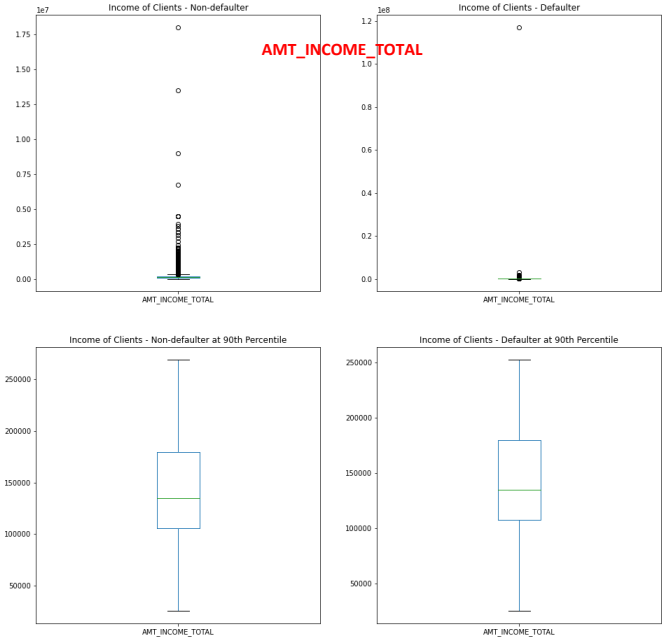


Continuation…

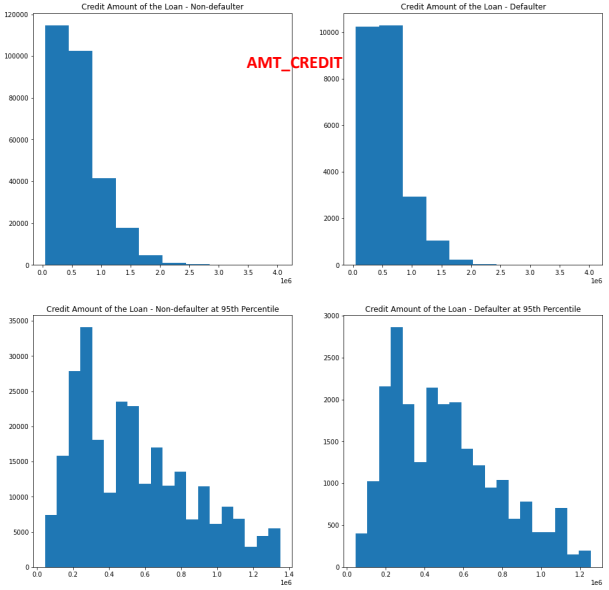
Univariate Analysis for Continuous Variables

Continuous Variables	Analysis
AMT_INCOME_TOTAL	The first two charts show the loan applicants' income level, and the following plots are after handling the outliers. The range of the non-defaulters is more comprehensive; however, the median is almost at the same level. Observations are based upon capping the data to the 90th percentile.
AMT_CREDIT	Non-defaulters tend to apply for higher loans as compared to non-defaulters.
AMT_ANNUITY	Non-defaulters tend to apply for higher loans as compared to non-defaulters.
AMT_GOODS_PRICE	The price of the goods on which the loans are spent was almost at the same level for both datasets.
DAYS_EMPLOYED	Applicants with more days of employment tend to be non-defaulters.

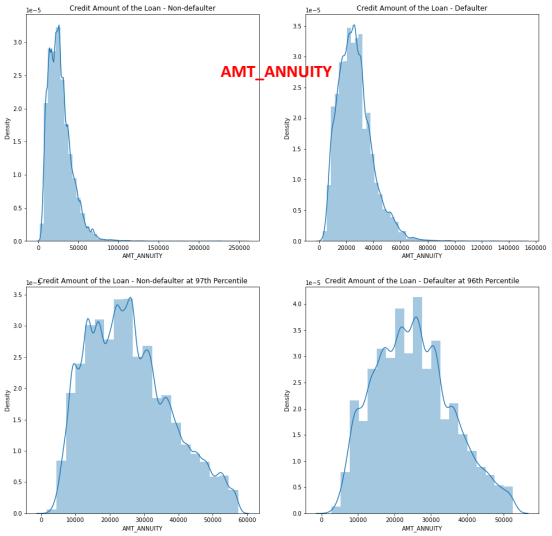
AMT_INCOME_TOTAL



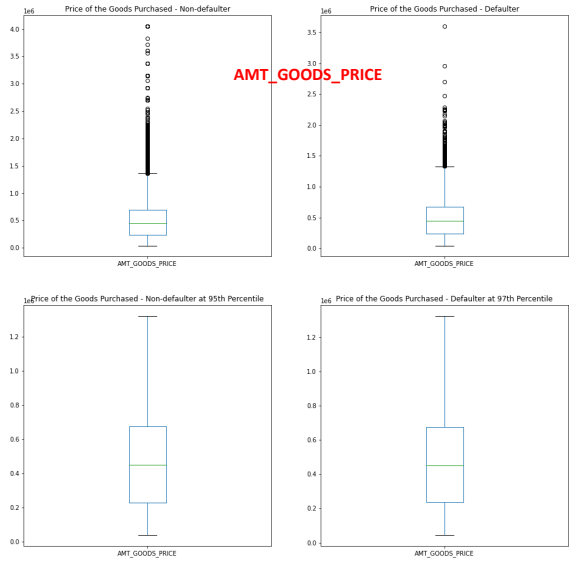
AMT_CREDIT



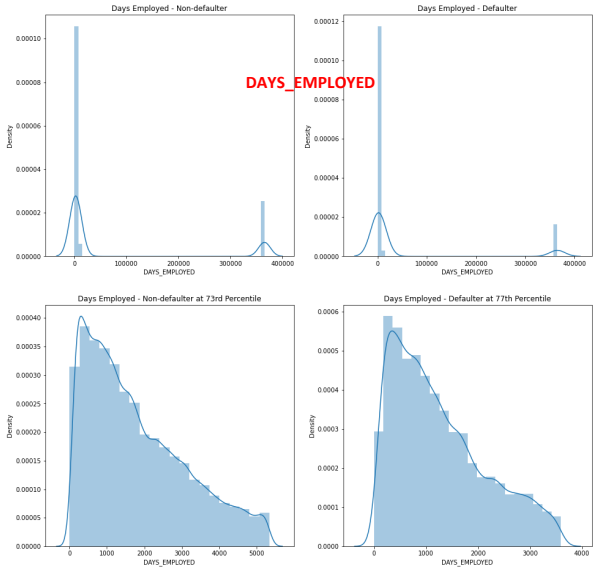
AMT_ANNUITY



AMT_GOODS_PRICE



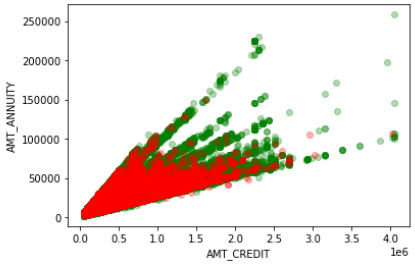
DAYS_EMPLOYED

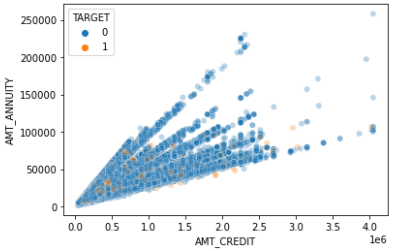


Continuation…

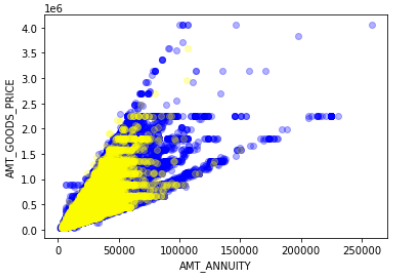
Bivariate Analysis Between Continuous Variables

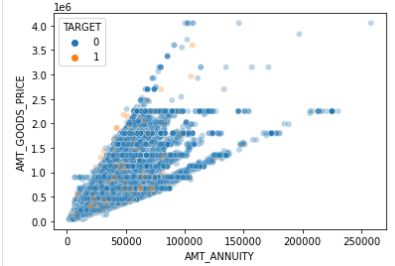
AMT_CREDIT vs AMT_ANNUITY
A high correlation between AMT_CREDIT and AMT_ANNUITY can be observed. There is a direct relationship in both variables depending on the amount of the bank's credit amount



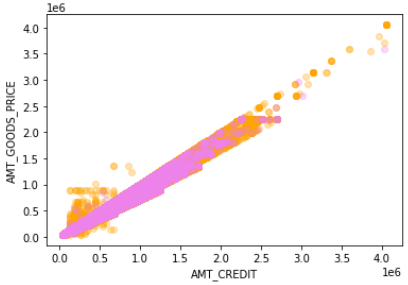


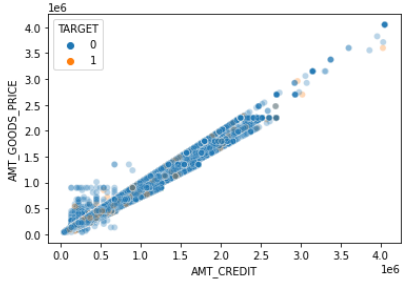
AMT_ANNUITY vs AMT_GOODS_PRICE
The exact high correlation is observed for AMT_ANNUITY and AMT_GOODS_PRICE. The higher the price of the purchased item will be higher the annuity due to higher credit.





AMT_CREDIT vs AMT_GOODS_PRICE
The highest correlation observed amongst the continuous variables is between AMT_CREDIT and AMT_GOODS_PRICE. This is expected as the credit is dependent on the price of the goods being purchased.





Continuation…

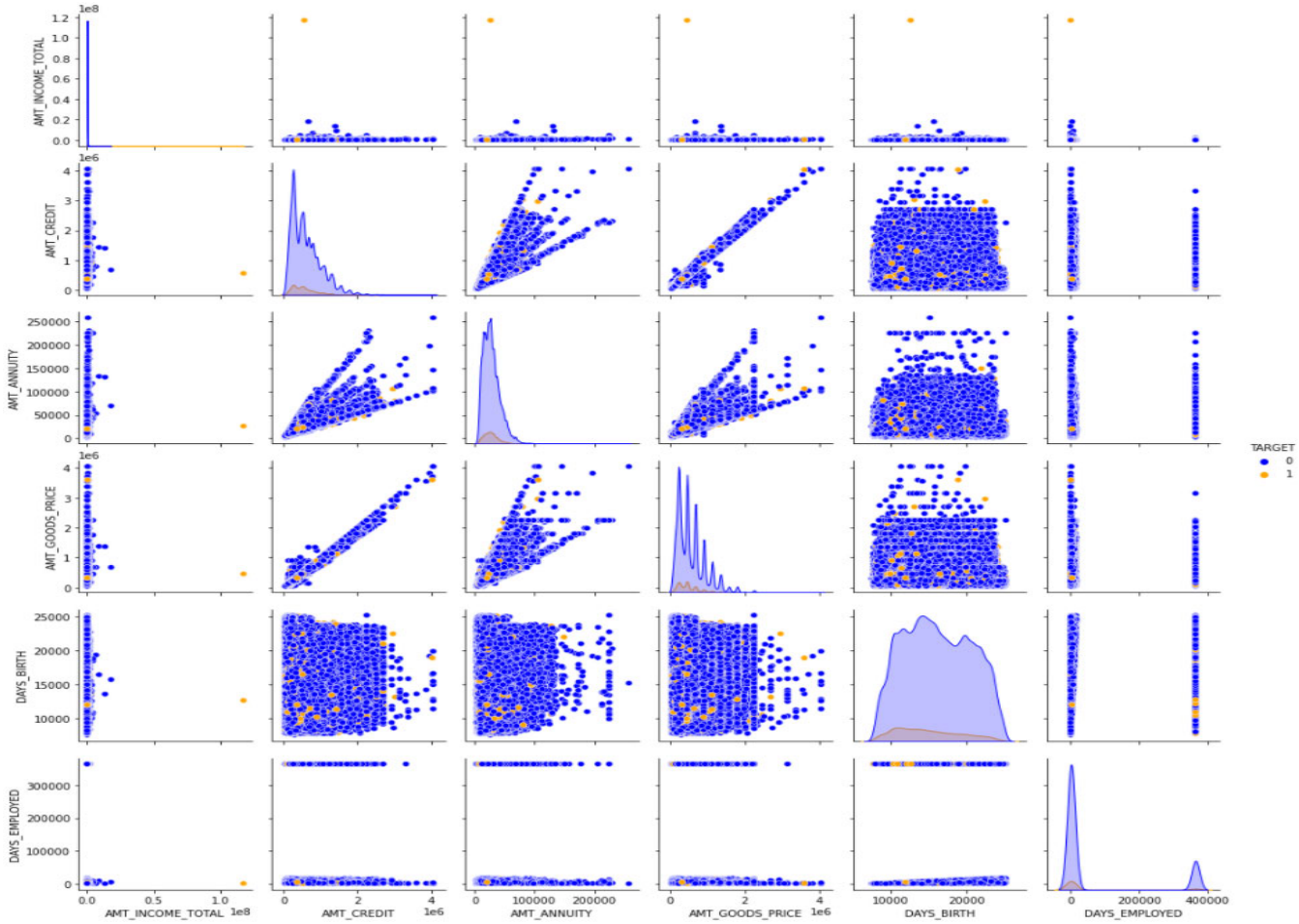
Bivariate Analysis Between Continuous Variables

- AMT_INCOME_TOTAL vs AMT_CREDIT vs AMT_ANNUITY vs AMT_GOODS_PRICE vs DAYS_BIRTH vs DAYS_EMPLOYED vs TARGET

One-stop visualization tool to see the correlation of each continuous variable.

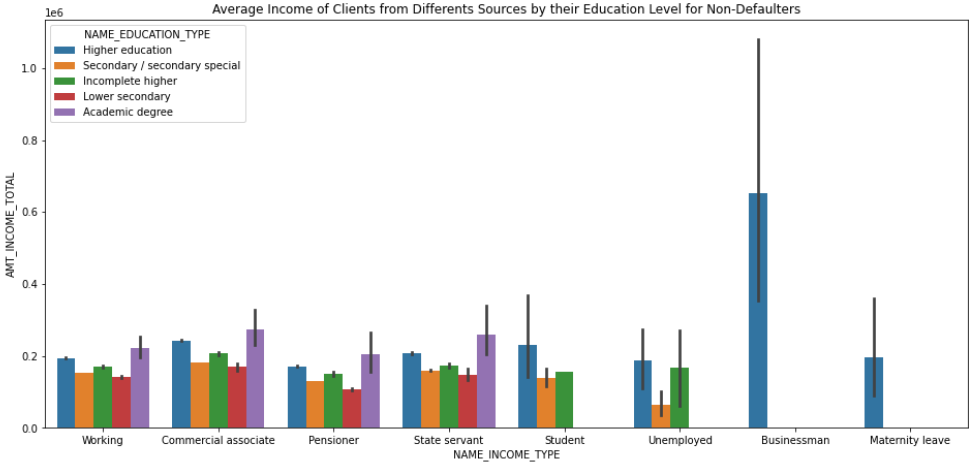
TARGET

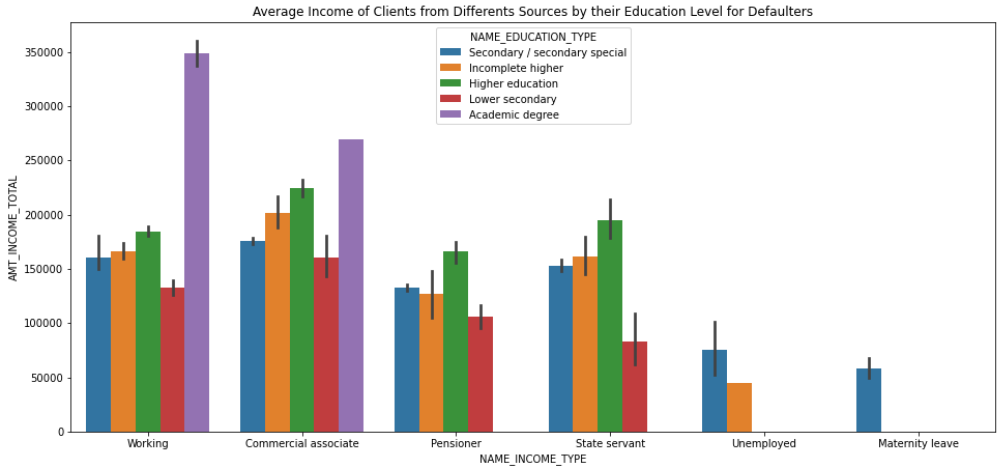
- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE
- DAYS_BIRTH
- DAYS_EMPLOYED



Continuation...

Bivariate/Multivariate Analysis Between Continuous and/or Categorical Variables





Non-Defaulters

Highest income comes from Businessman. For each income type, one of the major representatives comes from those who have attained Higher Education.

Defaulters

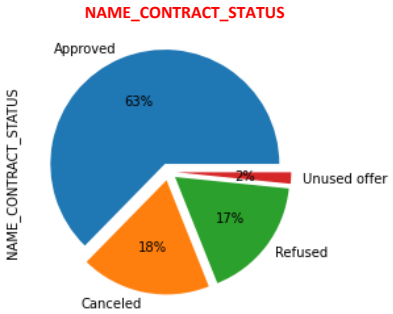
Income profile is at a lower end as compared to Non-Defaulters. No income type comes from Students and Businessmen.

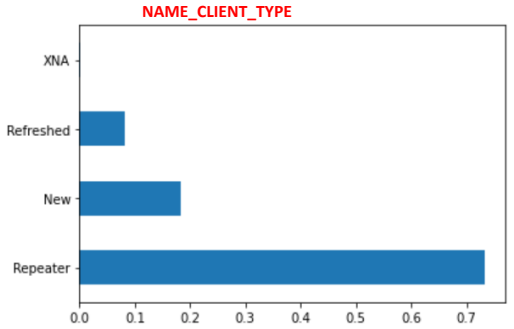
6. Data Visualization

Merging Application Data (ca) with Pervious Application (pa) Data

- **Inner** join on variable **SK_ID_CURR**

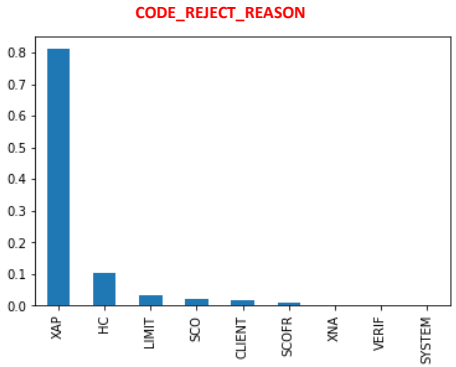
```
In [186]: # Merging the two datasets using inner join on SK_ID_CURR
          cpa = pd.merge(ca, pa, how = 'inner', on = 'SK_ID_CURR')

In [197]: cpa.shape
Out[197]: (1413701, 119)
```

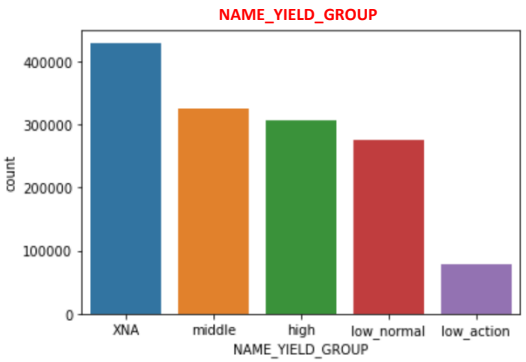


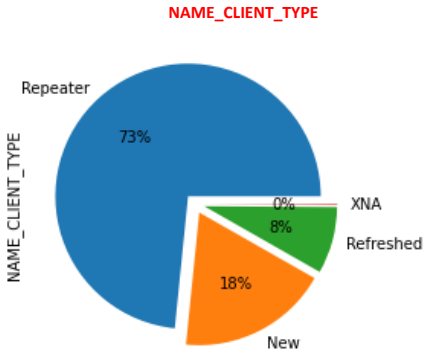


Univariate Analysis for the New Columns

Univariate variable	Analysis
NAME_CONTRACT_STATUS	It is showing the distribution of Contract Status for the previous loan applications of clients. The majority of the applications are approved.
NAME_CLIENT_TYPE	The majority of the loan applicants are repeat clients with a history of applying for a loan. XNA should have been removed as it is considered a null value.
NAME_YIELD_GROUP	XNA should have been removed as it is considered a null value. The majority of the interest rate is offered at the middle range, while the lowest interest has the lowest frequency.
CODE_REJECT_REASON	XAP should have been removed as it is considered a null value. The lowest occurrence of rejection of loan applications is due to the bank system and verification processes.
NAME_CLIENT_TYPE	XAP should have been removed as it is considered a null value. The majority of the loan applicants are repeat customers.



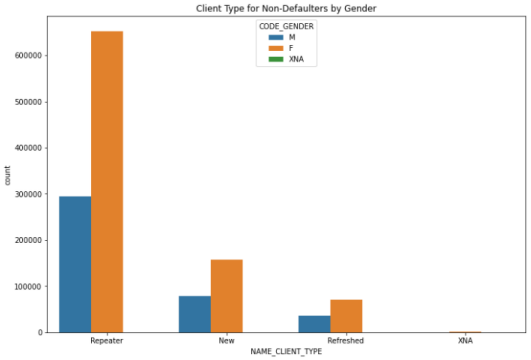


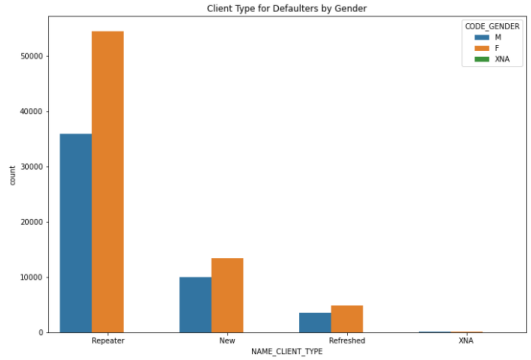


Continuation…

NAME_CLIENT_TYPE vs CODE_GENDER

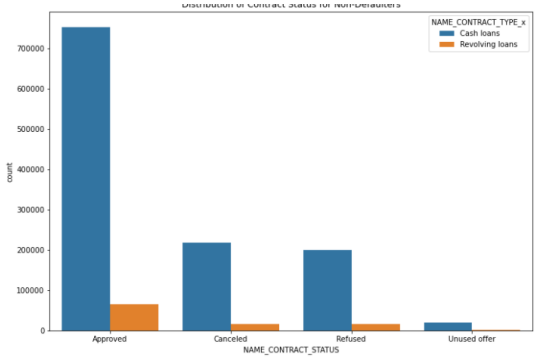
Female repeat and new applicants tend to default less than males.

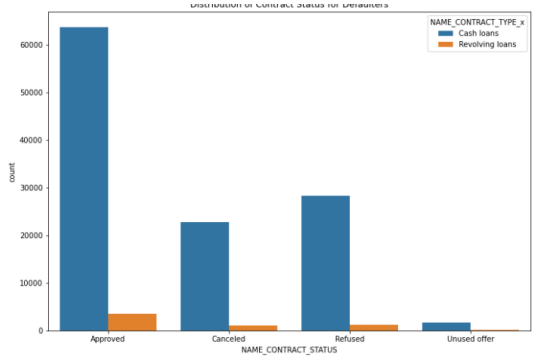




NAME_CONTRACT_STATUS vs NAME_CONTRACT_TYPE_x

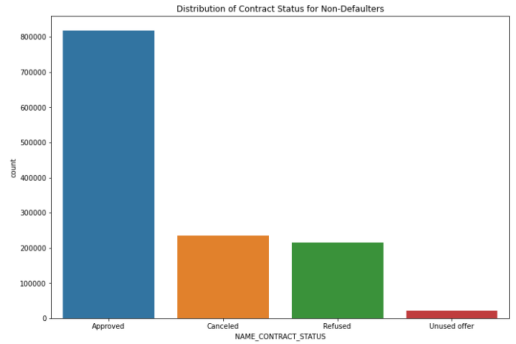
Cash loans of defaulter have a higher rate of refusal.

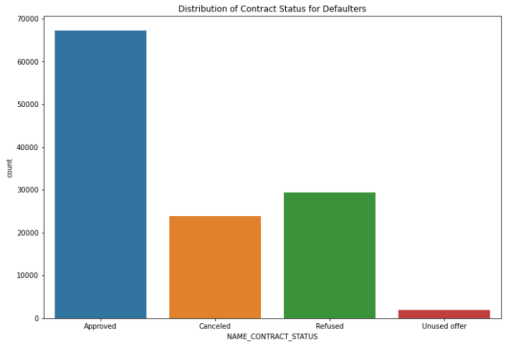




NAME_CONTRACT_STATUS

It is observed that defaulters have a higher refusal rate.





Continuation…

Correlation of All Variables for the Combined Dataset

	Var1	Var2	Corr
6494	AMT_GOODS_PRICE_y	AMT_APPLICATION	1.00
1153	FLAG_EMP_PHONE	DAYS_EMPLOYED	1.00
3469	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
3111	FLOORSMAX_MEDI	FLOORSMAX_AVG	1.00
3022	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.99
6495	AMT_GOODS_PRICE_y	AMT_CREDIT_y	0.99
3113	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.99
532	AMT_GOODS_PRICE_x	AMT_CREDIT_x	0.99
2935	FLOORSMAX_MODE	FLOORSMAX_AVG	0.99
6318	AMT_CREDIT_y	AMT_APPLICATION	0.98

Non - Defaulter

	Var1	Var2	Corr
6494	AMT_GOODS_PRICE_y	AMT_APPLICATION	1.00
1153	FLAG_EMP_PHONE	DAYS_EMPLOYED	1.00
3469	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
3111	FLOORSMAX_MEDI	FLOORSMAX_AVG	1.00
3022	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.99
6495	AMT_GOODS_PRICE_y	AMT_CREDIT_y	0.99
3113	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.99
532	AMT_GOODS_PRICE_x	AMT_CREDIT_x	0.99
2935	FLOORSMAX_MODE	FLOORSMAX_AVG	0.99
6318	AMT_CREDIT_y	AMT_APPLICATION	0.98

Defaulter

	Var1	Var2	Corr
1153	FLAG_EMP_PHONE	DAYS_EMPLOYED	1.00
6494	AMT_GOODS_PRICE_y	AMT_APPLICATION	1.00
3469	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
3022	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	1.00
3111	FLOORSMAX_MEDI	FLOORSMAX_AVG	1.00
6495	AMT_GOODS_PRICE_y	AMT_CREDIT_y	0.99
3113	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.99
2935	FLOORSMAX_MODE	FLOORSMAX_AVG	0.99
2846	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.98
532	AMT_GOODS_PRICE_x	AMT_CREDIT_x	0.98

It would be difficult for now to know the real variables that can be considered as highly correlated to TARGET variable as no imputation and outlier handling has been done to the final dataset (cpa). The only columns dropped are those from the current application data (ca) with null values >50%. The variables with highest correlation could have been visually identified by plotting a heatmap using relevant columns/variables only. The irrelevant columns could have been dropped decreasing further significantly the total variables being considered in the analysis.

7. Recommendation

We analyzed almost 150+ variables and looked into various factors based on the stats and visual graphs. Results look very competitive with a thin margin; however, we identified some key variables that could potentially impact the bank in protecting credit or interest loss. You could find a recommendation below that could contribute to the approval process customer loan as reference parameters.

- Working professionals are more likely to apply loans, and more significant loans are from secondary educational attainment levels.
- Stats show more focus should be on cash loans as compared to other kinds of loans.
- Owned car applicants do not prefer to take a loan; however, those who have owned car have their residence, which means this is recommended to focus on non-owned car applicant.
- The volume of applicants staying in an apartment is more prominent in number; however, this does not signify to approve or reject the loan.
- Widows and separated are not motivated towards the loan application, and subsequently, the chances are high that they will be a defaulter.
- Those with less or no children are more motivated towards are the loan application, and they are non-defaulters.
- Business entities are more align with the loan and provide a high yield of interest.
- Laborers, sales staff, Core staff, Managers, Drivers are among the top applicants for a loan; however, no significant variables tell these may or may not defaulters.
- Females customers apply for more loans, and under the repeater category, they are more non-defaulter.
- For defaulter history, cash loans pop up more on the rejection queue.
- It can be observed that the loan annuity of the defaulters is higher than those who are non-defaulters.
- Applicants with more days of employment tend to be non-defaulters.
- The highest income comes from Businessman. For each income type, one of the significant representatives comes from those who have attained Higher Education.
- Income profile is at a lower end as compared to Non-Defaulters. No income type comes from Students and Businessmen.

End of the case Study, Thank you ☺

Name:- **Mohammad Shahid Rashid**
Mohammad.shahid.rashid@gmail.com
Batch :- DCS27

Name:- **Kirk Xyrrus Villanueva**
kirkxyrrusvillanueva@yahoo.com.ph
Batch :- DCS27