# Credit EDA Case Study

Presented by –

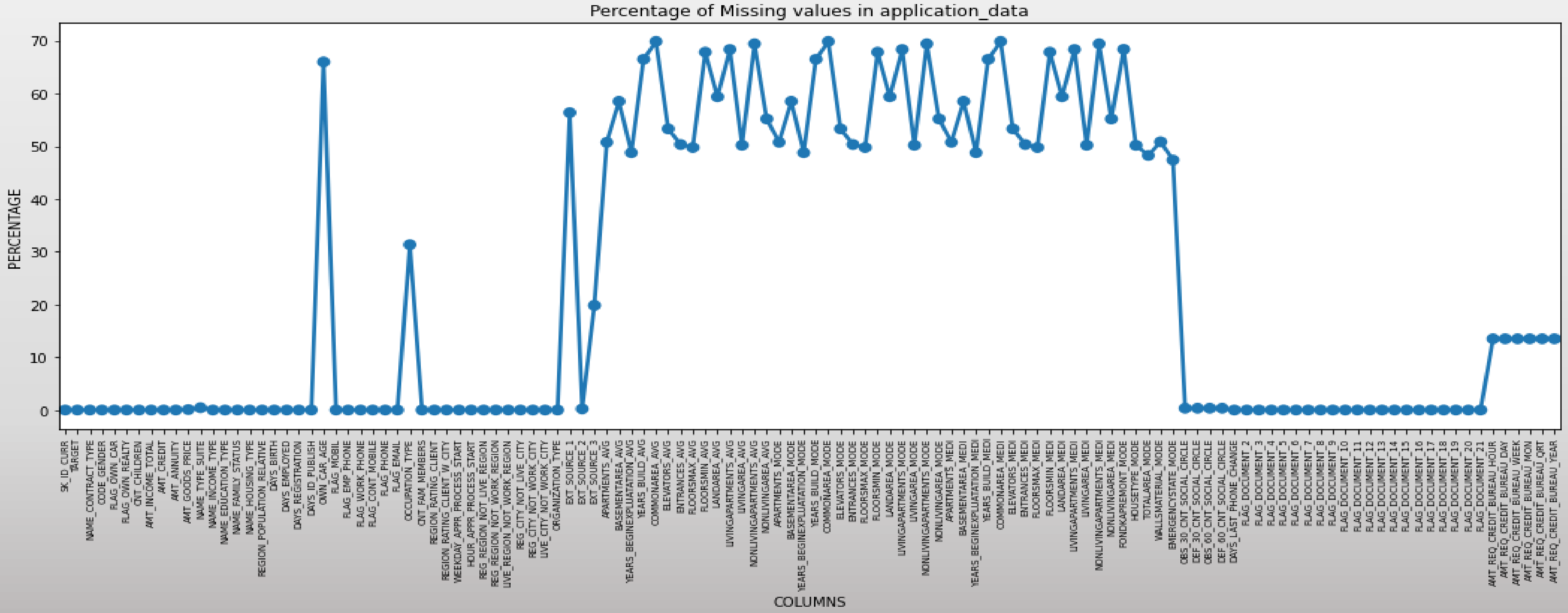
Moonmoon Sinha

Oiendrila Nath

DS C34

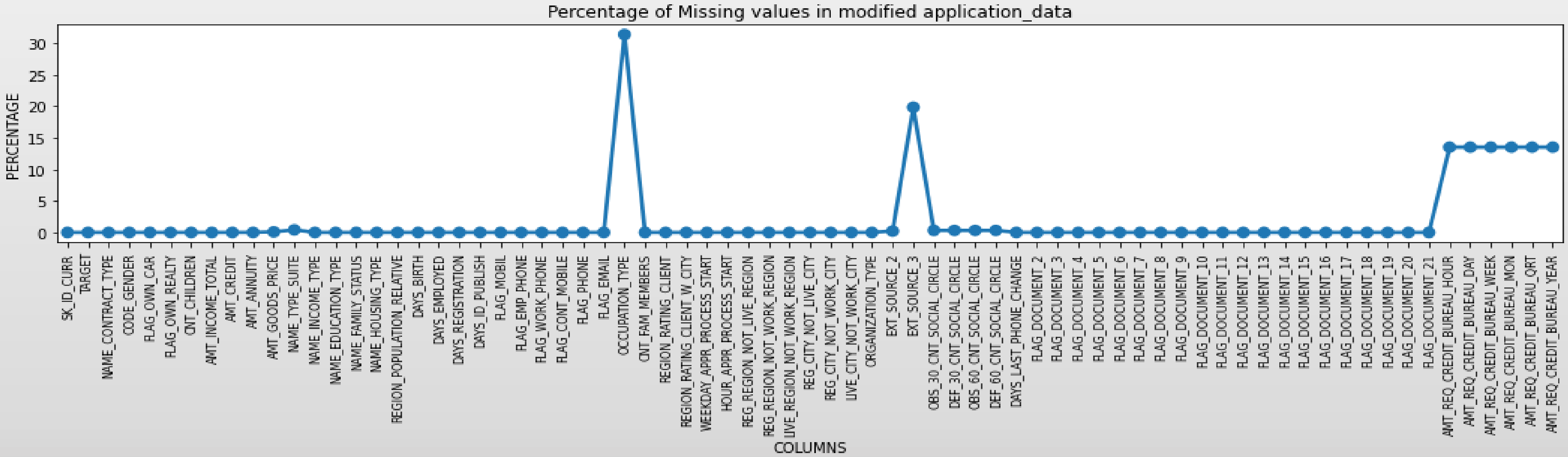
# Finding Missing Values in Current Application Dataset

- The plot below shows all the missing values in current application file :



# Dropping Missing Values in Current Application Dataset

- Columns having more than 45% missing values in application file have been dropped as shown in the plot below. The dropped columns majorly described the building/ apartment/ common area where the loan applicants reside.



- The dataset initially had 307511 rows and 122 columns. After dropping the missing value columns, there are 307511 rows and 73 columns remaining in the dataset.

# Missing Value Treatment in Current Application Dataset

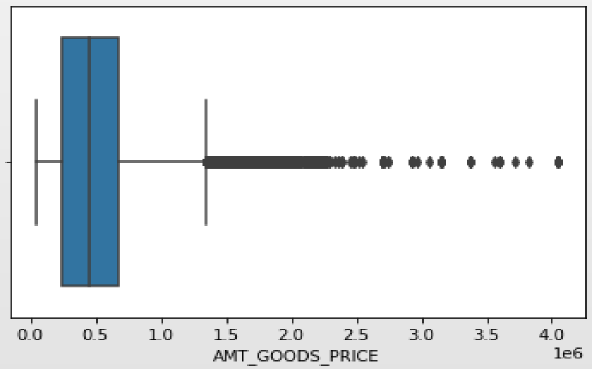
- Next we have to impute the missing values in the columns that have lower missing value counts
- Columns that need missing value treatment :

| COLUMN INDEX |
|---|
| OCCUPATION_TYPE |
| EXT_SOURCE_3 |
| 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 |
| NAME_TYPE_SUITE |
| OBS_30_CNT_SOCIAL_CIRCLE |
| DEF_30_CNT_SOCIAL_CIRCLE |
| OBS_60_CNT_SOCIAL_CIRCLE |
| DEF_60_CNT_SOCIAL_CIRCLE |
| EXT_SOURCE_2 |
| AMT_GOODS_PRICE |

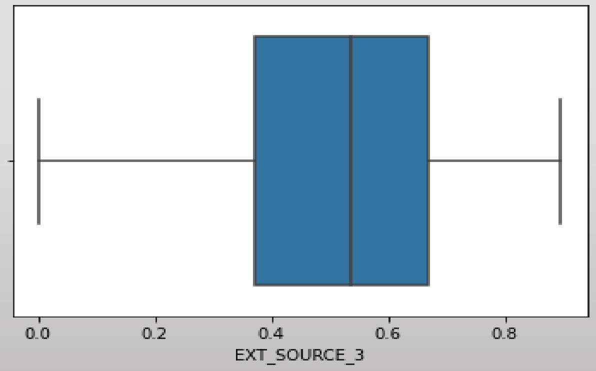
- The negative values in days columns have been changed to absolute values.

# Missing Value Treatment in Numerical Columns

- Imputing missing values with median in numerical data columns having outliers :

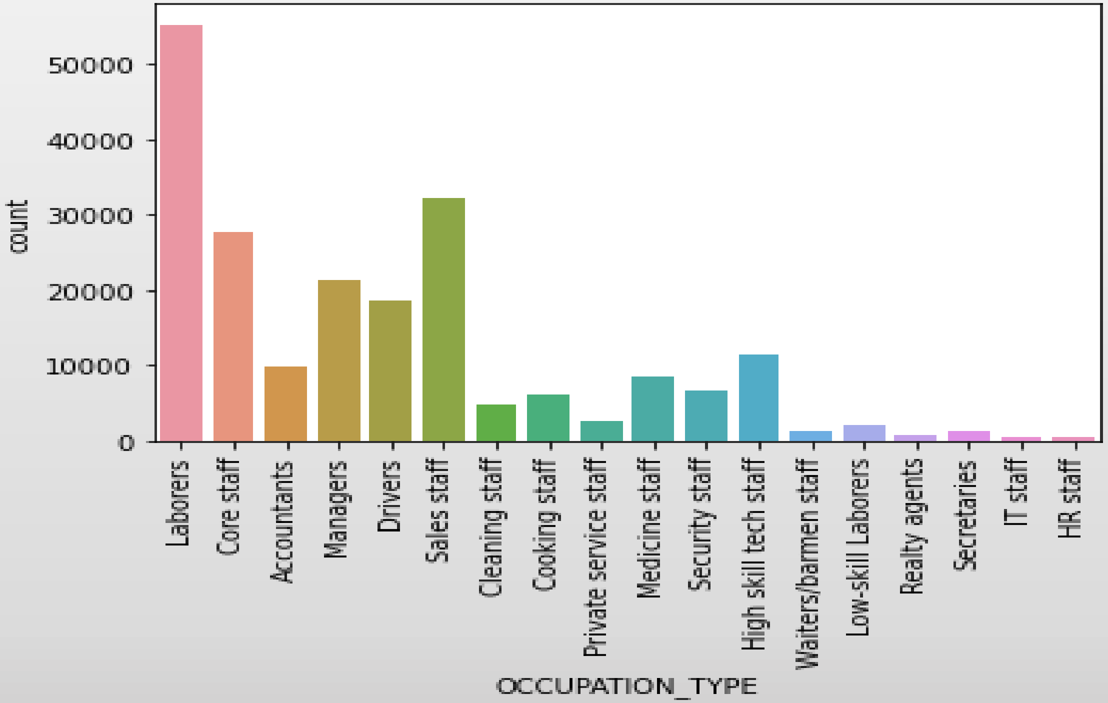


- Imputing missing values with mean in numerical data columns not having outliers :



# Missing Value Treatment in Categorical Columns

- Imputing missing values with mode in categorical columns :



- It can be observed that the maximum number of loan applications were received from the labourers.

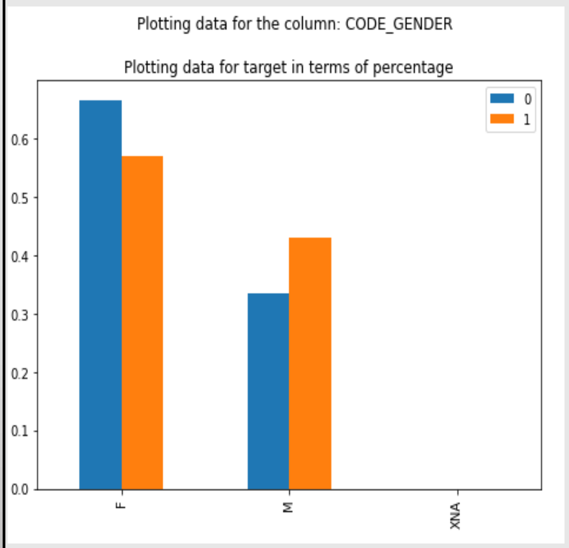
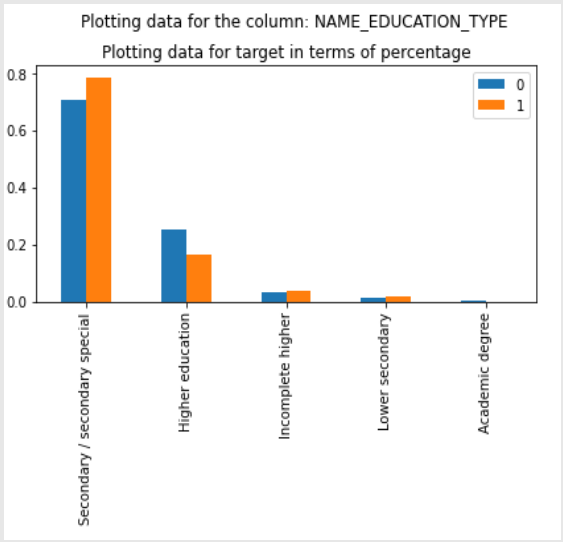
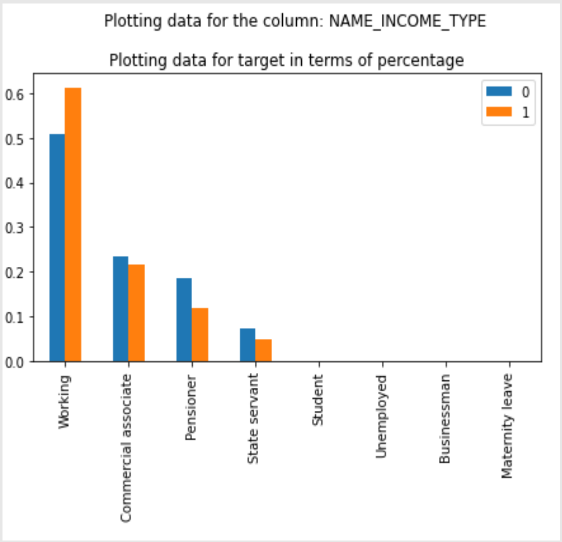
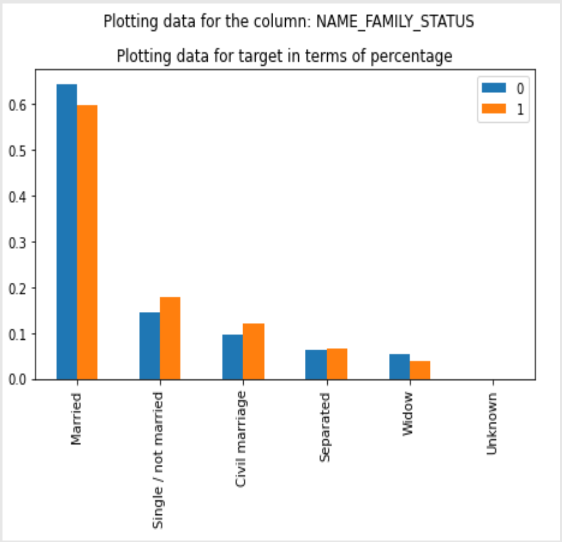
# Analyzing the Target

- The target column gives the insight on the clients with payment difficulties (target = 1) and other clients (target = 0).

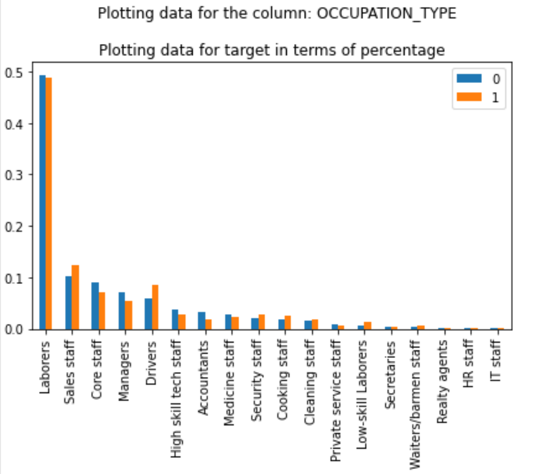
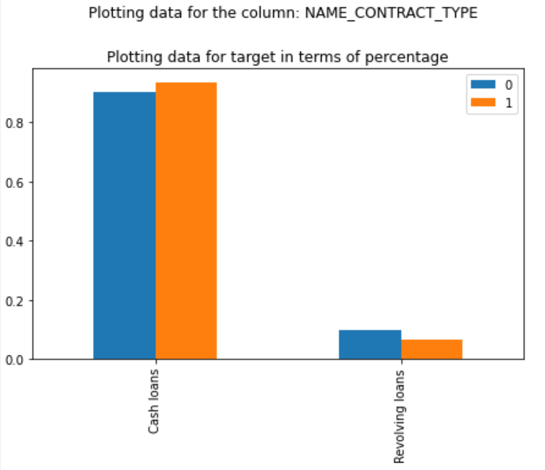
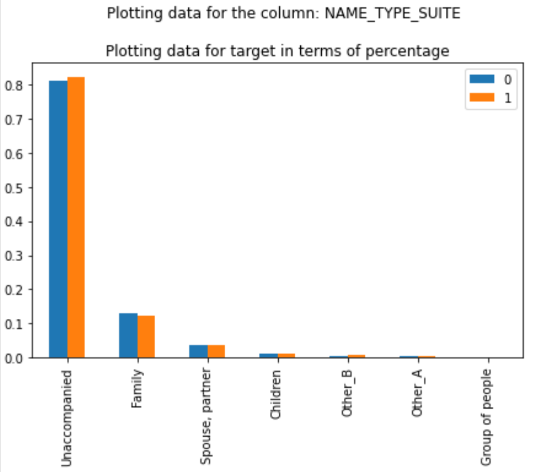
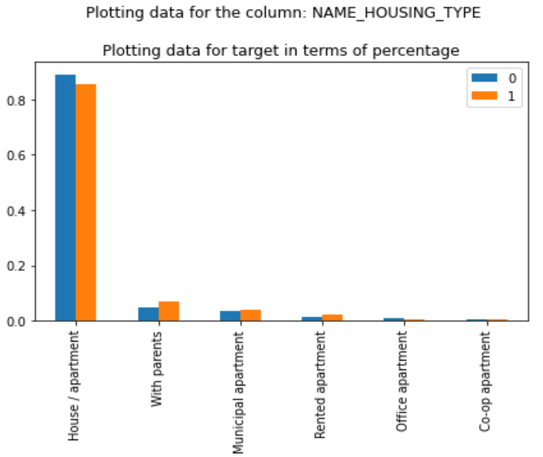
- On analysis, it can be seen that out of the total 307511 clients, 24825 clients have payment difficulties, i.e. 8.07% clients have payment difficulties.

| TARGET | Percentage (%) |
|---|---|
| 0 | 91.93 |
| 1 | 8.07 |

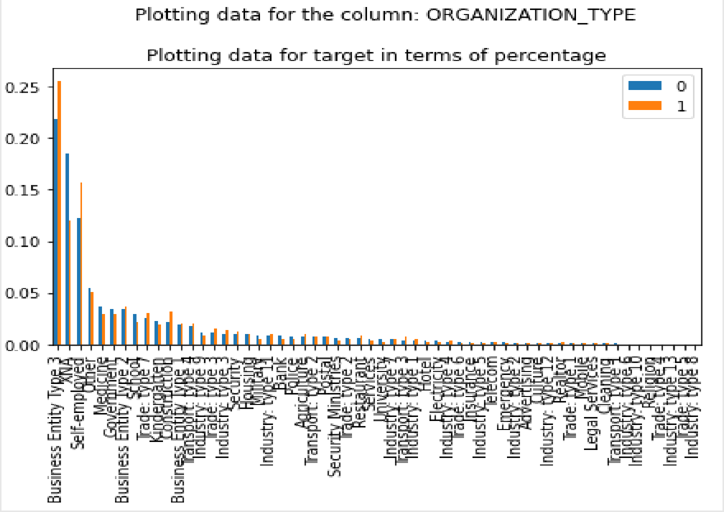
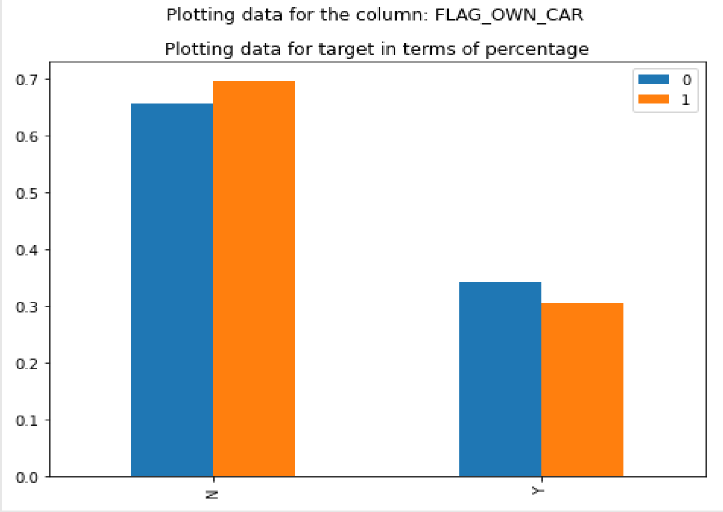
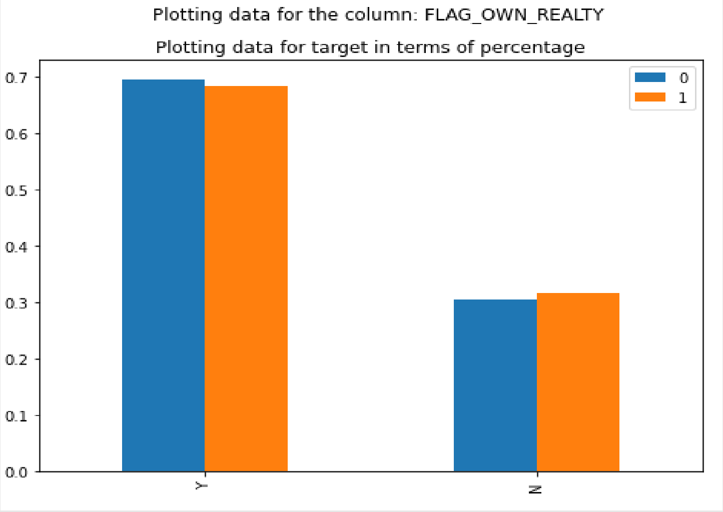
# Univariate Analysis of Target

| Observations | | | |
|---|---|---|---|
| Less number of males take loans but the defaulters are higher in case of males. | Most customers take loan for secondary education followed by higher education. But the default rate in secondary education is much high and for higher education is comparatively low. | Although working professionals are the major loan applicants, they have a high default rate compared to the pensioners. | Married people are more likely to opt for loans while they are more likely to be the defaulters. Notably, single/ civil married/ separated candidates are highly likely to be defaulters. |
|  |  |  |  |

# Univariate Analysis of Target

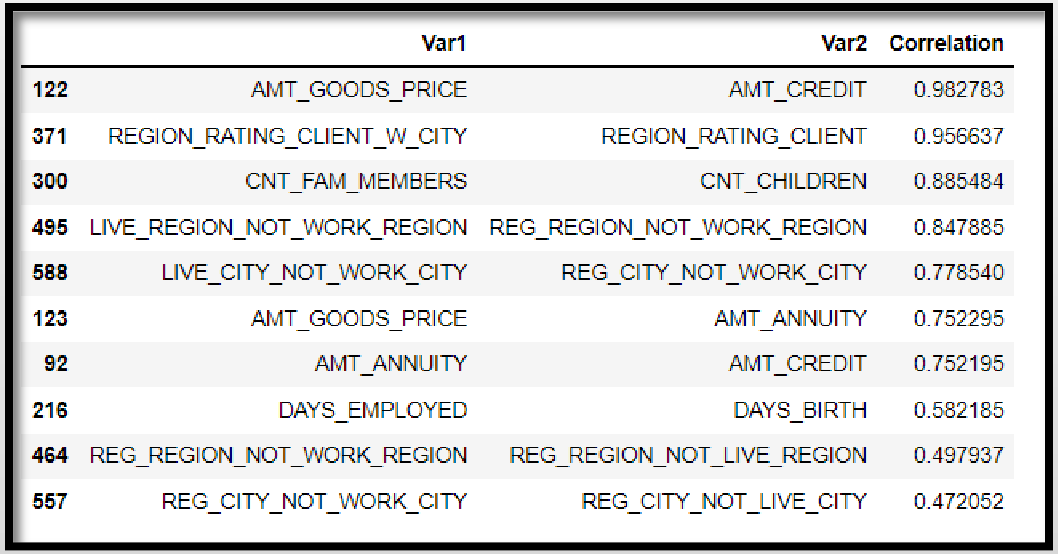
<table>
<tr><th colspan="4">Observations</th></tr>
<tr>
<td>Clients with low income are more probable to take loans and become defaulters.</td>
<td>Clients more likely opt for cash loans than revolving loans while the defaulters are also greater in case of cash loans.</td>
<td>Major loan clients who were unaccompanied, took loans and later had difficulties in payment.</td>
<td>Clients living in house/ apartment more often take loans but clients who live with parents or in municipal or rented apartments are more likely to be defaulters.</td>
</tr>
<tr>
<td>
Plotting data for the column: OCCUPATION_TYPE
Plotting data for target in terms of percentage
</td>
<td>
Plotting data for the column: NAME_CONTRACT_TYPE
Plotting data for target in terms of percentage
</td>
<td>
Plotting data for the column: NAME_TYPE_SUITE
Plotting data for target in terms of percentage
</td>
<td>
Plotting data for the column: NAME_HOUSING_TYPE
Plotting data for target in terms of percentage
</td>
</tr>
</table>

# Univariate Analysis of Target

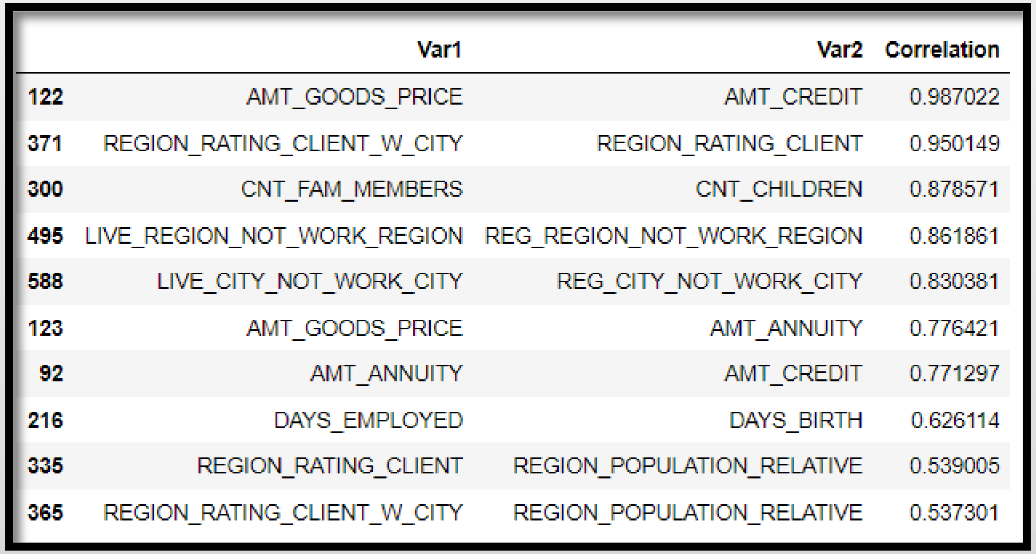
| Observations | | |
|---|---|---|
| Clients belonging to business entity type 3, self-employed and XNA mostly apply for loans while the percentage of defaulters is found to be higher in the first two cases. | Clients who do not own cars are more likely to take loans. Notably, the percentage of defaulters is also higher in these cases. | Clients who own a house or flat take more loans, while the clients who do not own a house or flat are more likely to be defaulters. |
|  |  |  |

# Bivariate Analysis of Target

- Top 10 correlations of the current application dataset have been recorded in the below tables both in case of defaulters and non-defaulters.



| | Var1 | Var2 | Correlation |
|---|---|---|---|
| 122 | AMT_GOODS_PRICE | AMT_CREDIT | 0.982783 |
| 371 | REGION_RATING_CLIENT_W_CITY | REGION_RATING_CLIENT | 0.956637 |
| 300 | CNT_FAM_MEMBERS | CNT_CHILDREN | 0.885484 |
| 495 | LIVE_REGION_NOT_WORK_REGION | REG_REGION_NOT_WORK_REGION | 0.847885 |
| 588 | LIVE_CITY_NOT_WORK_CITY | REG_CITY_NOT_WORK_CITY | 0.778540 |
| 123 | AMT_GOODS_PRICE | AMT_ANNUITY | 0.752295 |
| 92 | AMT_ANNUITY | AMT_CREDIT | 0.752195 |
| 216 | DAYS_EMPLOYED | DAYS_BIRTH | 0.582185 |
| 464 | REG_REGION_NOT_WORK_REGION | REG_REGION_NOT_LIVE_REGION | 0.497937 |
| 557 | REG_CITY_NOT_WORK_CITY | REG_CITY_NOT_LIVE_CITY | 0.472052 |



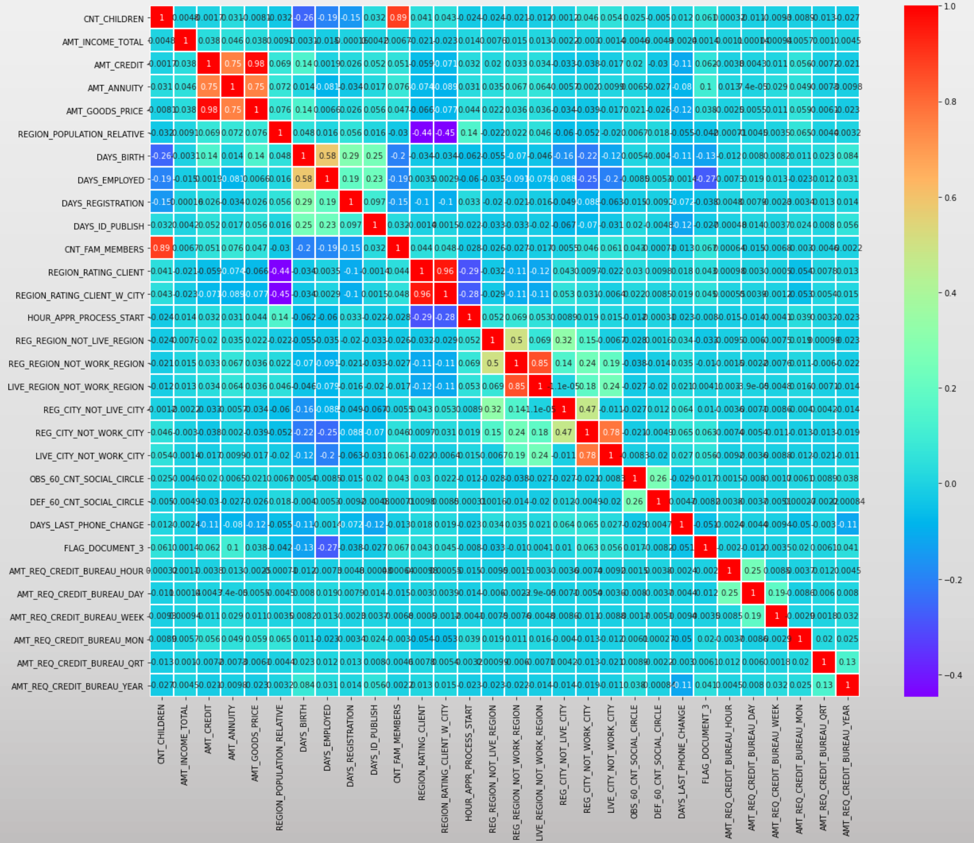
| | Var1 | Var2 | Correlation |
|---|---|---|---|
| 122 | AMT_GOODS_PRICE | AMT_CREDIT | 0.987022 |
| 371 | REGION_RATING_CLIENT_W_CITY | REGION_RATING_CLIENT | 0.950149 |
| 300 | CNT_FAM_MEMBERS | CNT_CHILDREN | 0.878571 |
| 495 | LIVE_REGION_NOT_WORK_REGION | REG_REGION_NOT_WORK_REGION | 0.861861 |
| 588 | LIVE_CITY_NOT_WORK_CITY | REG_CITY_NOT_WORK_CITY | 0.830381 |
| 123 | AMT_GOODS_PRICE | AMT_ANNUITY | 0.776421 |
| 92 | AMT_ANNUITY | AMT_CREDIT | 0.771297 |
| 216 | DAYS_EMPLOYED | DAYS_BIRTH | 0.626114 |
| 335 | REGION_RATING_CLIENT | REGION_POPULATION_RELATIVE | 0.539005 |
| 365 | REGION_RATING_CLIENT_W_CITY | REGION_POPULATION_RELATIVE | 0.537301 |

Clients with payment difficulties

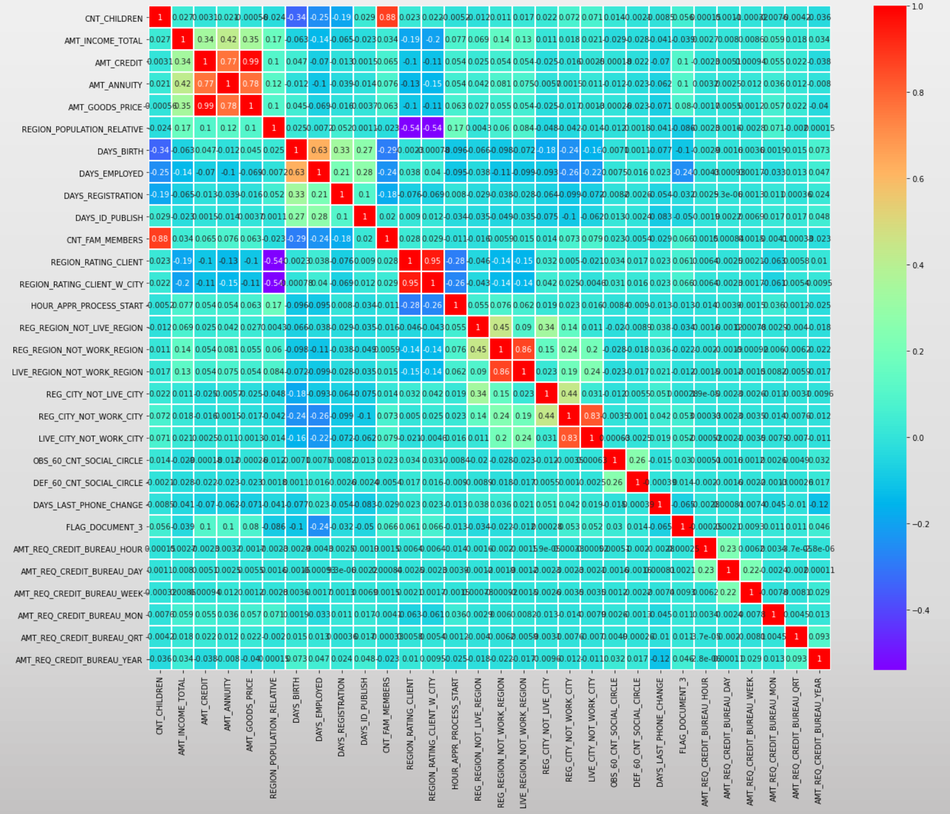
Clients without payment difficulties

# Bivariate Analysis of Target

## Heatmaps showing correlation between different variables



Clients with payment difficulties



Clients without payment difficulties

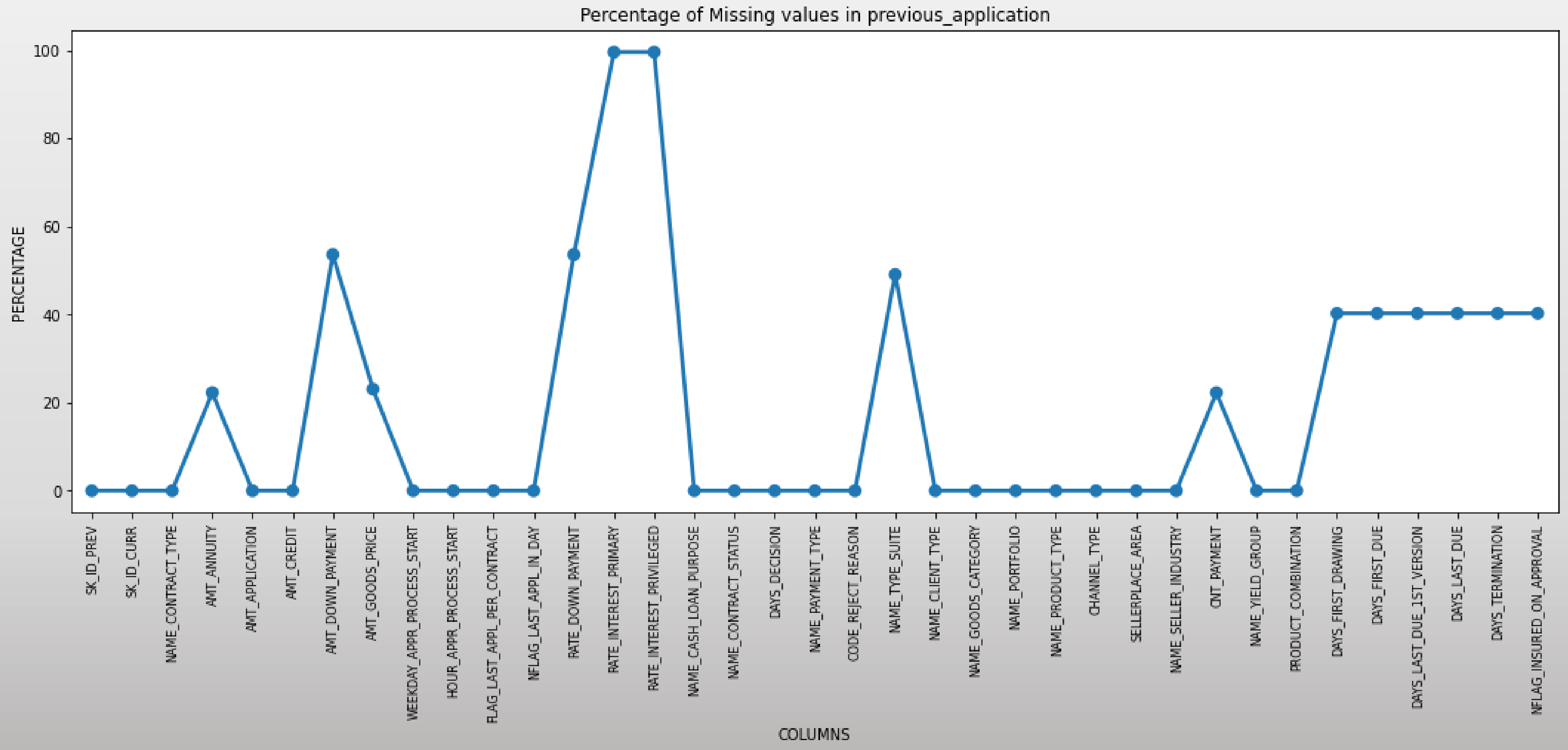
# Bivariate Analysis of Target

Observations from the heatmap :

- Among the top 10 correlations, the goods price amount and credit amount are highly correlated in case of both defaulters and non-defaulters.
- The annuity amount and credit amount are also highly correlated although the correlation is slightly less in defaulters (75%) compared to non-defaulters (77%).
- The correlation in the number of days employed and current age of the client is high in non-defaulters (62%) compared to defaulters (58%).

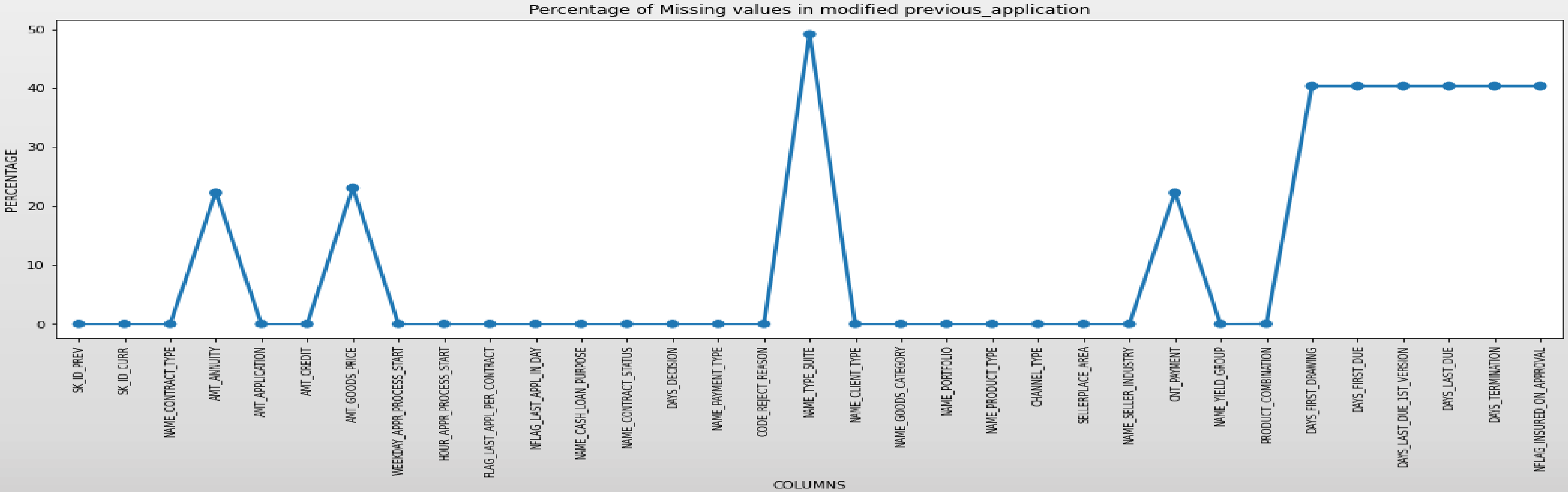
# Finding Missing Values in Previous Application Dataset

- The plot below shows all the missing values in current application file :



# Dropping Missing Values in Previous Application Dataset

- Columns having more than 50% missing values in the previous application file have been dropped as shown in the plot below.



- The dataset initially had 1670214 rows and 37 columns. After dropping the missing value columns, there are 1670214 rows and 33 columns remaining in the dataset.

# Missing Value Treatment in Previous Application Dataset

- Next we have to impute the missing values in the columns that have lower missing value counts.
- Columns that need missing value treatment :

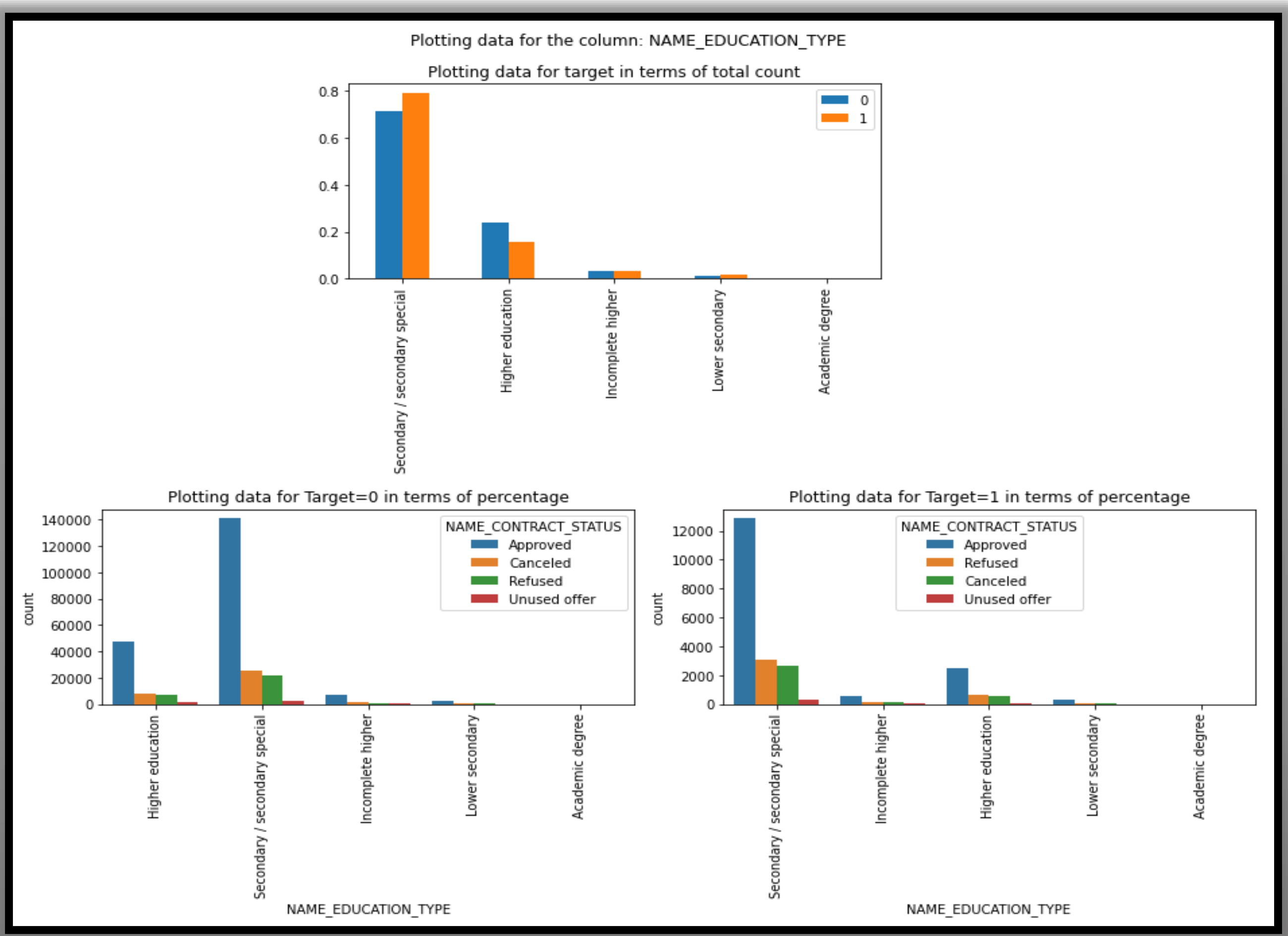
| COLUMN INDEX |
|---|
| AMT_ANNUITY |
| AMT_GOODS_PRICE |
| CNT_PAYMENT |
| AMT_CREDIT |

- We will not impute the missing values in the days columns.
- The negative values in days columns have been changed to absolute values.

# Univariate Analysis of Target in Merged Dataset

- Clients with secondary and higher education mostly opt for loans.

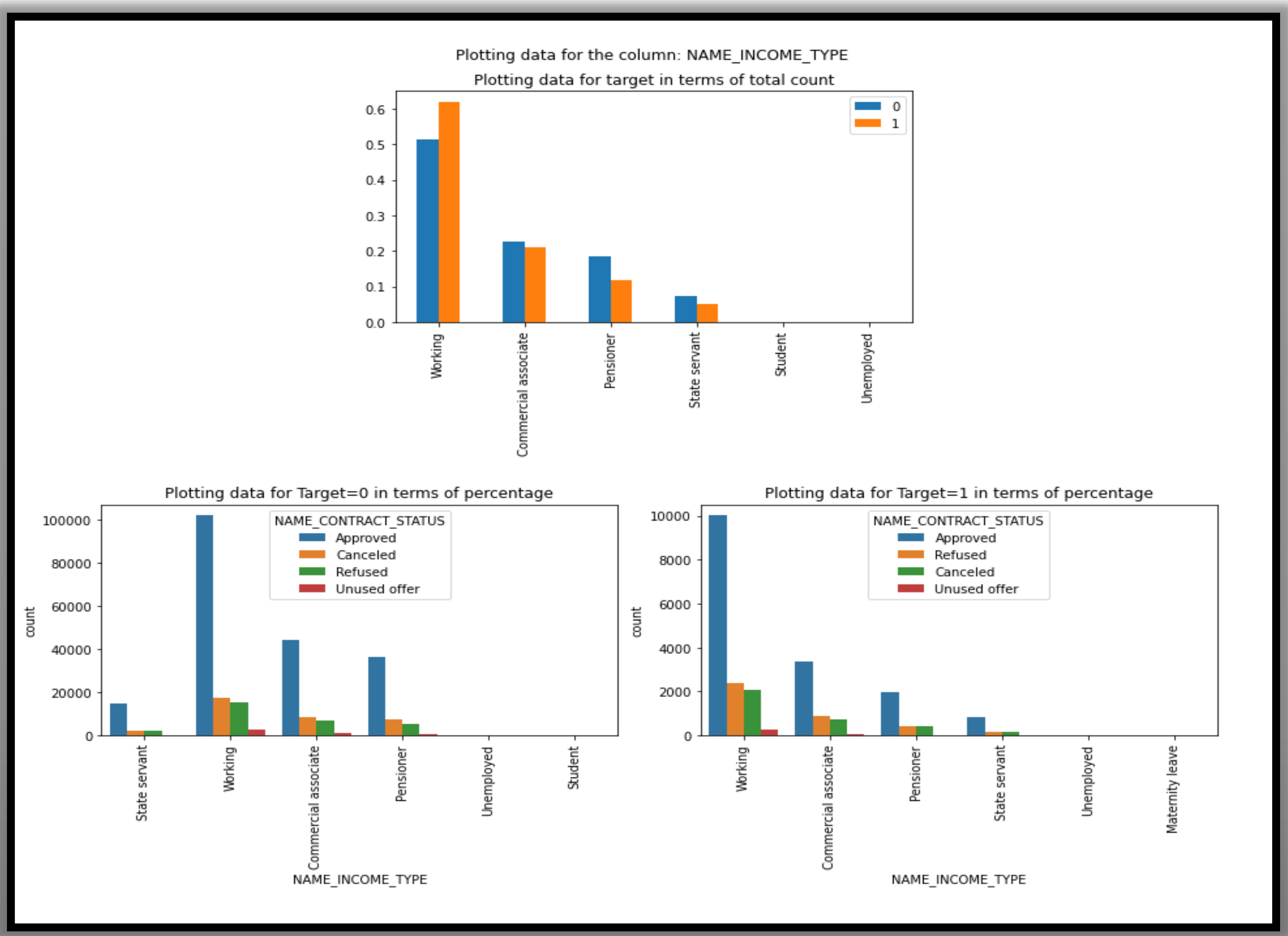
- Maximum number of secondary educated and higher educated clients with approved loans are defaulters.



# Univariate Analysis of Target in Merged Dataset

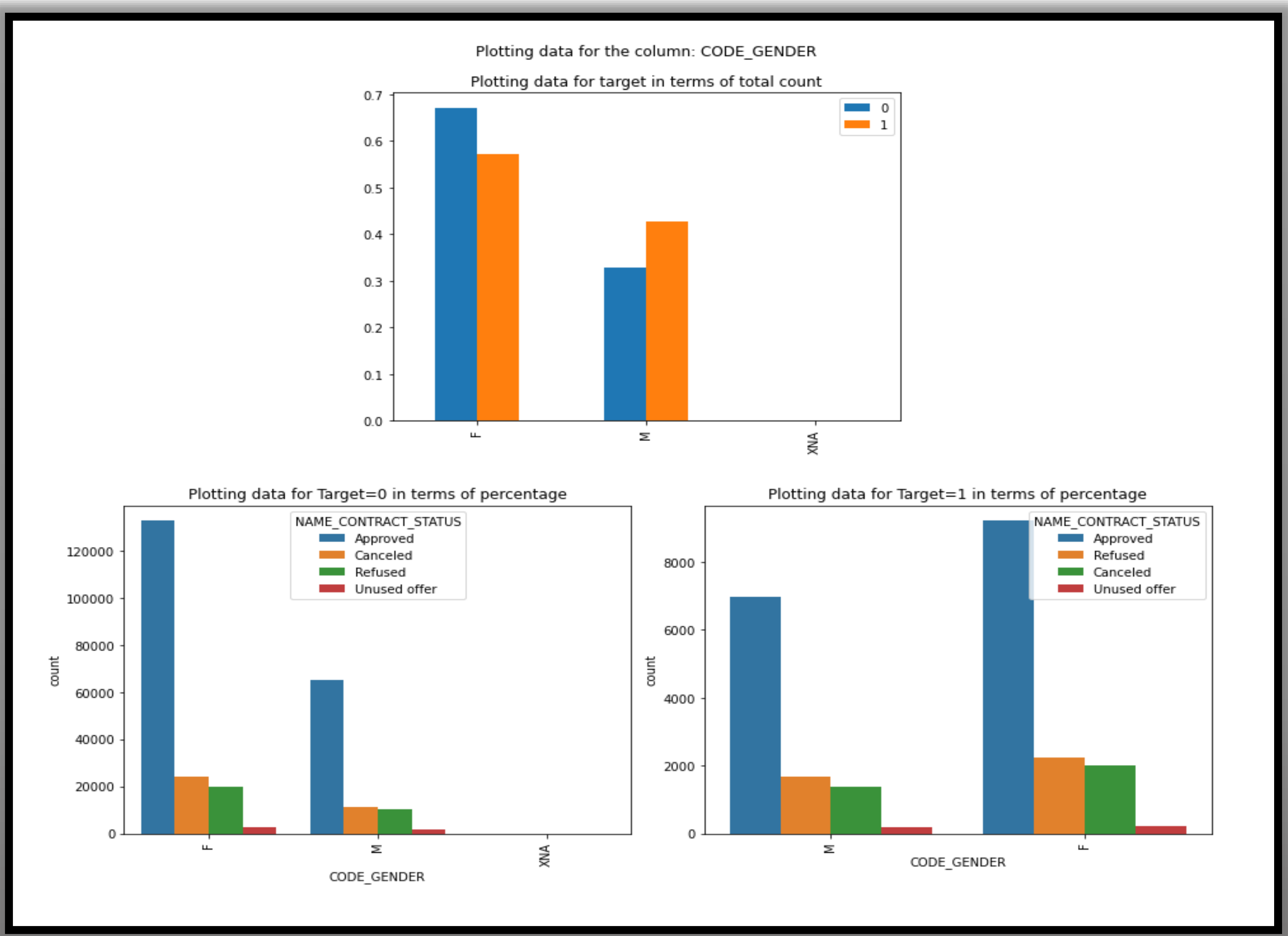
- Working professionals, pensioners and commercial associates mostly get loans approved.

- Pensioners are less likely to be defaulters as compared to working professionals.



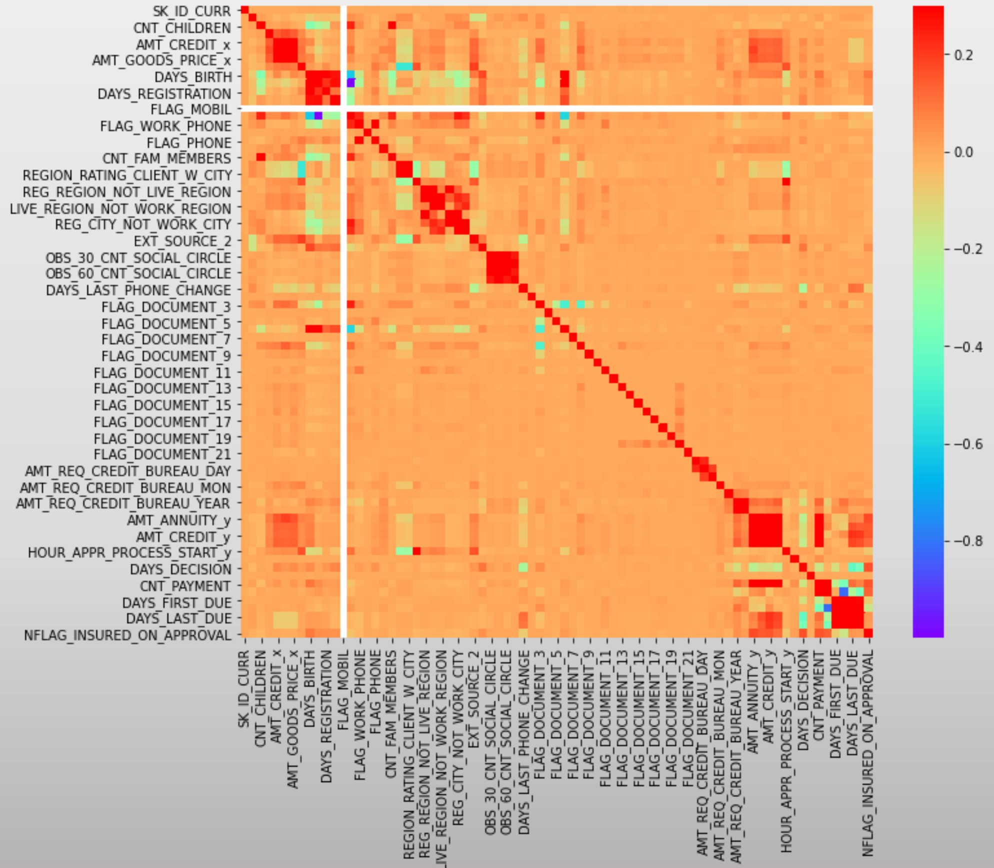
# Univariate Analysis of Target in Merged Dataset

- Female clients more frequently opt for loans compared to males.
- More male clients are defaulters compared to female clients.

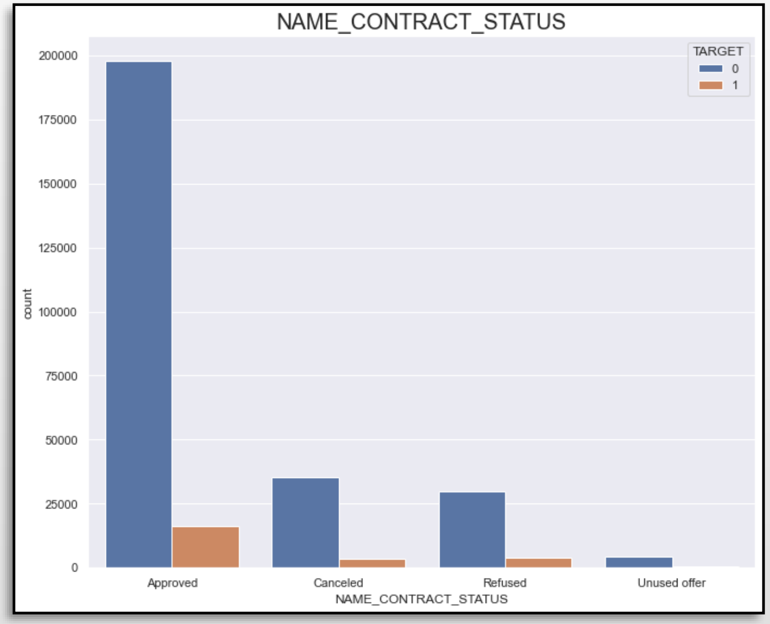


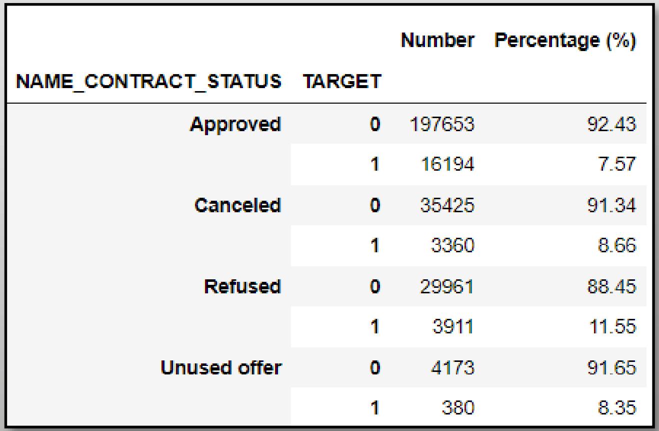
# Bivariate Analysis of Merged Dataset

- The below heatmap shows the correlation between the variables of the merged dataset. It is evident that AMT_CREDIT has high positive correlation with AMT_ANNUITY and AMT_GOODS_PRICE.



# Analysis Inferences





| NAME_CONTRACT_STATUS | TARGET | Number | Percentage (%) |
|---|---|---|---|
| Approved | 0 | 197653 | 92.43 |
| | 1 | 16194 | 7.57 |
| Canceled | 0 | 35425 | 91.34 |
| | 1 | 3360 | 8.66 |
| Refused | 0 | 29961 | 88.45 |
| | 1 | 3911 | 11.55 |
| Unused offer | 0 | 4173 | 91.65 |
| | 1 | 380 | 8.35 |

- Approximately 91% of the previously cancelled clients are actually non-defaulters.
  - Maybe they received worse pricing and if this is revised, these clients can be interested to take loans, thus leading to business growth.
- Clients who have been previously refused loans, have actually repaid them in approximately 88% cases.
  - Reasons for why these clients were refused loans should be reconsidered as they are less unlikely to be defaulters.
- About 91% of the clients who have unused the offer were highly likely to repay the loan.
  - Interest rates can be reviewed and if possible revised, as these clients would possibly not be defaulters.