

Credit EDA Case Study

PRESENTED BY –

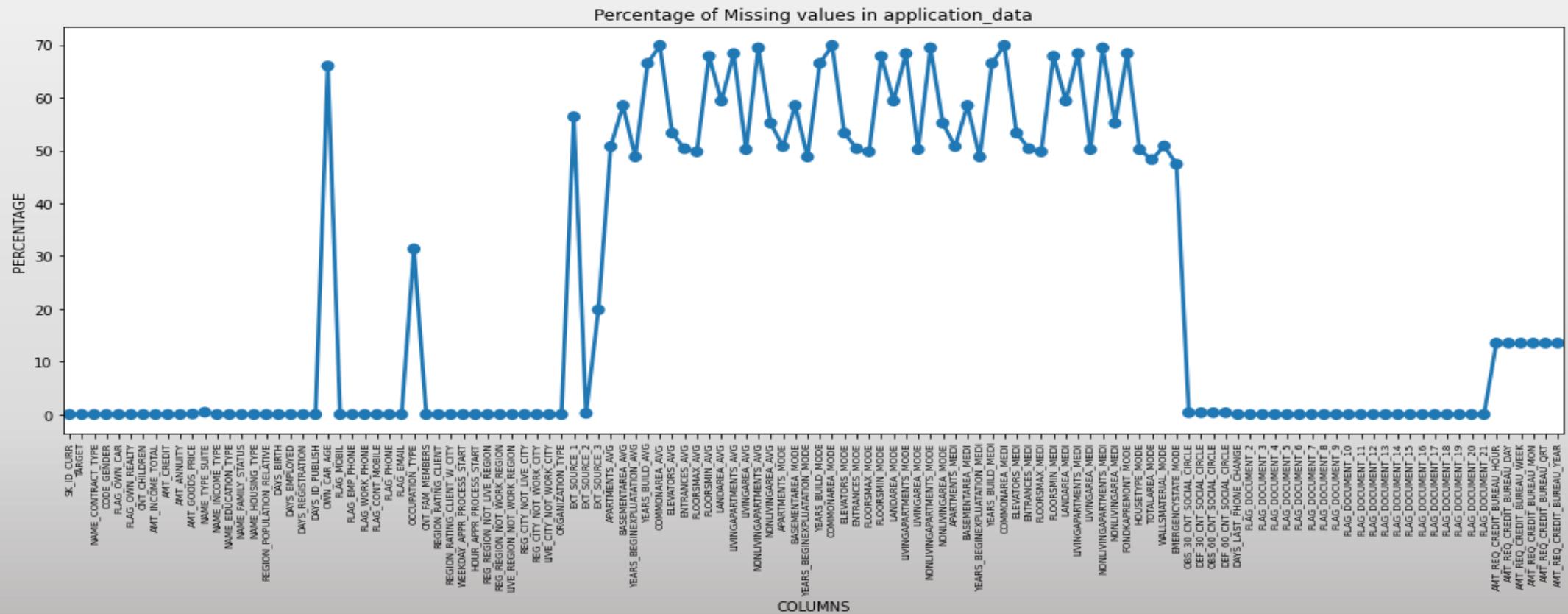
MOONMOON SINHA

OIENDRILA NATH

DS C34

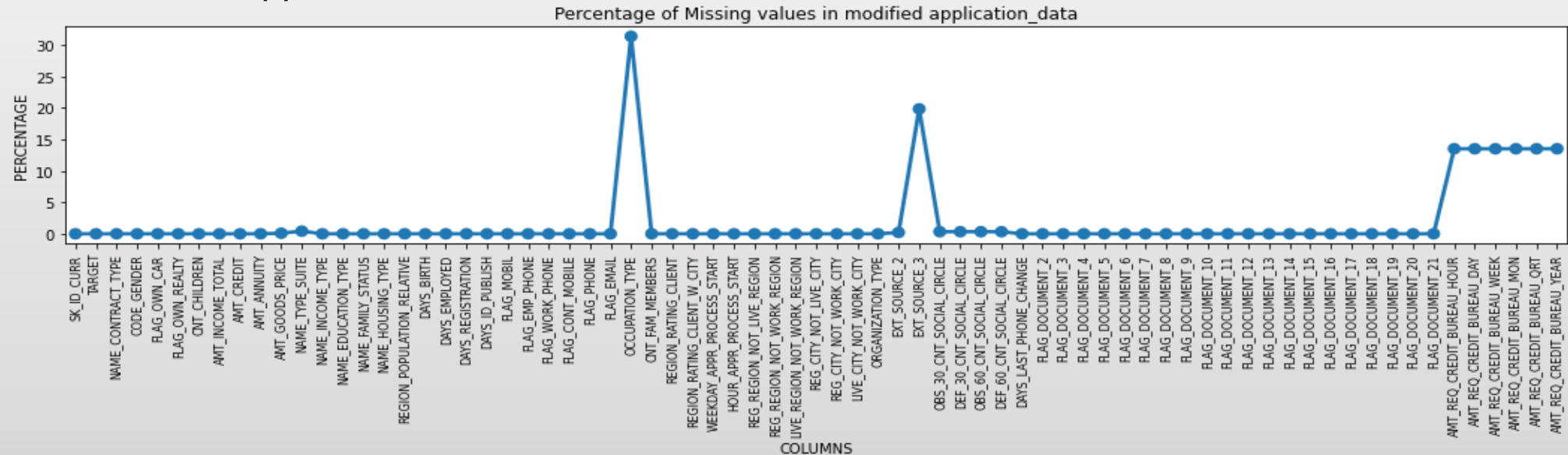
Finding Missing Values in Current Application Dataset

- 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

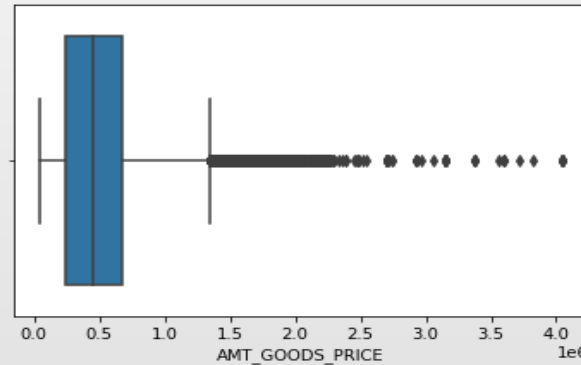
- Next we have to impute the missing values in the columns that have lower missing value counts
- Columns that need missing value treatment :

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

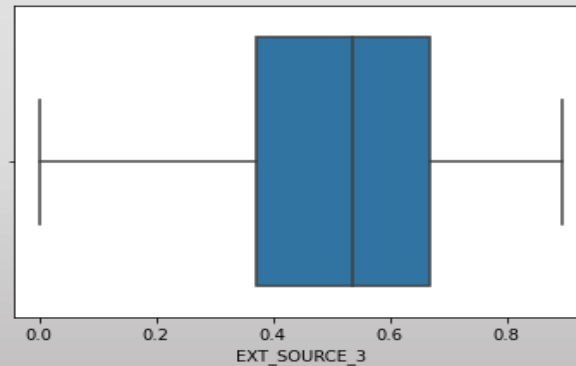
- The negative values in days columns have been changed to absolute values.

Missing Value Treatment in Numerical Columns

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

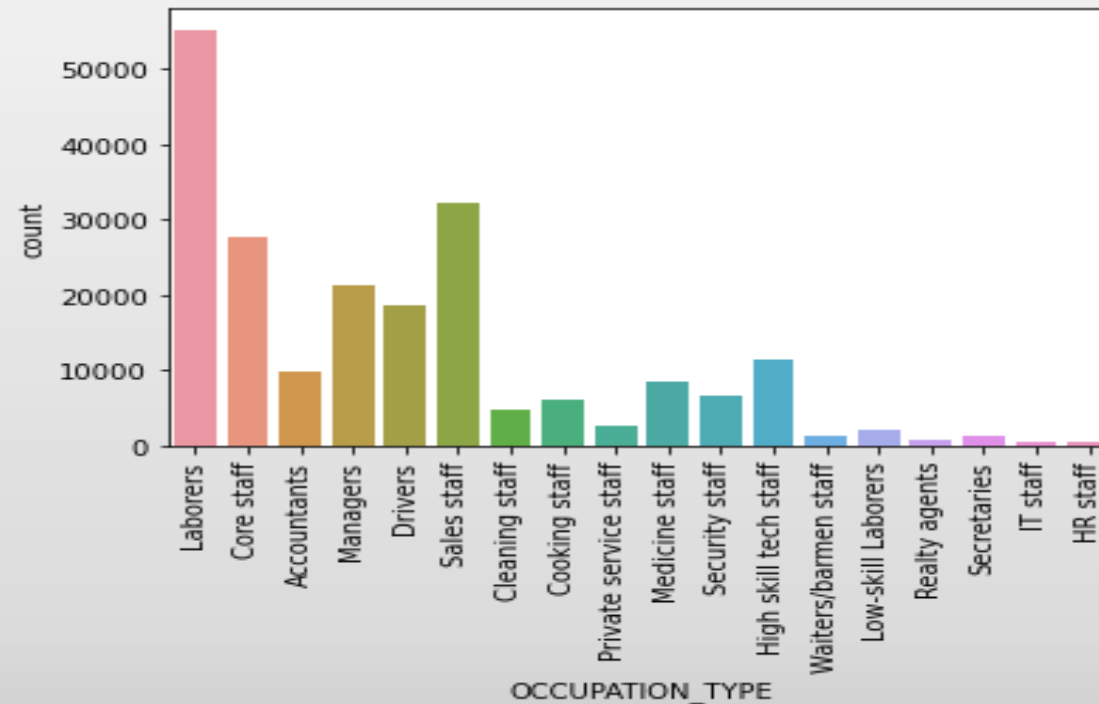


- Imputing missing values with mean in numerical data columns not having outliers :



Missing Value Treatment in Categorical Columns

- Imputing missing values with mode in categorical columns :



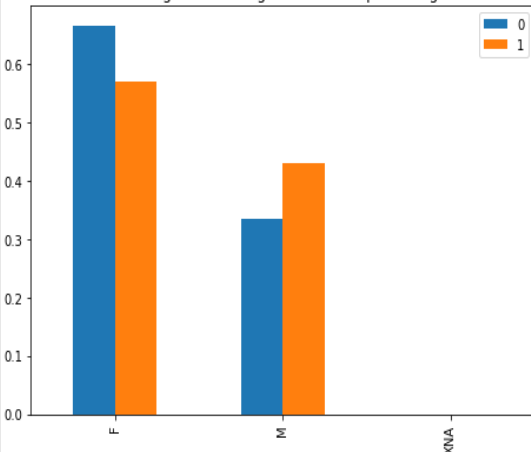
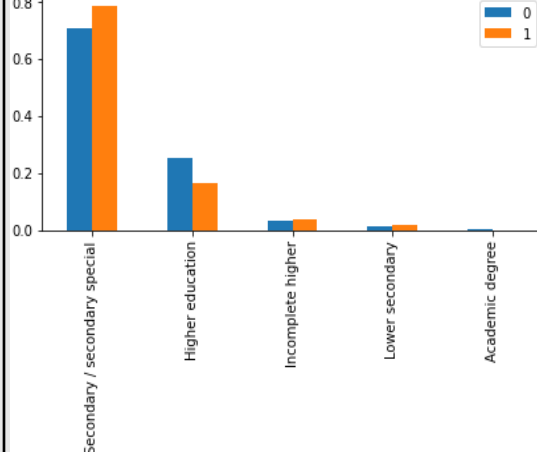
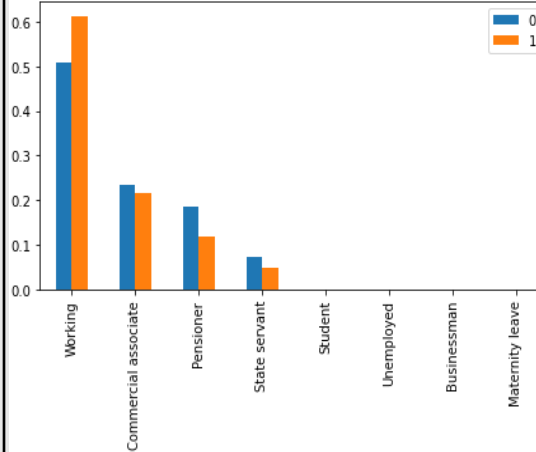
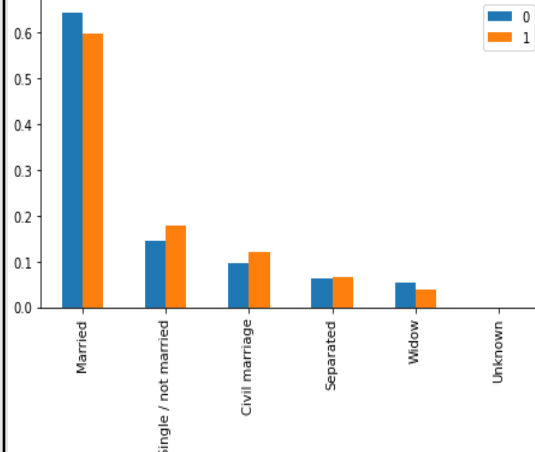
- It can be observed that the maximum number of loan applications were received from the labourers.

Analyzing the Target

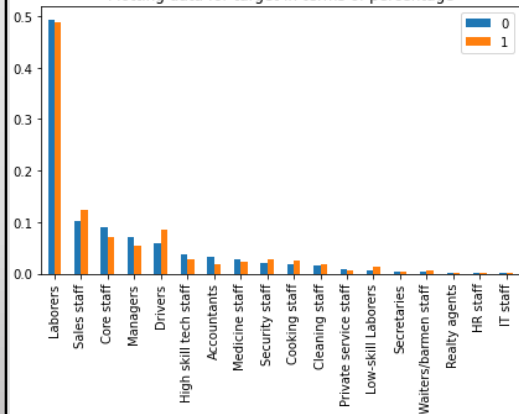
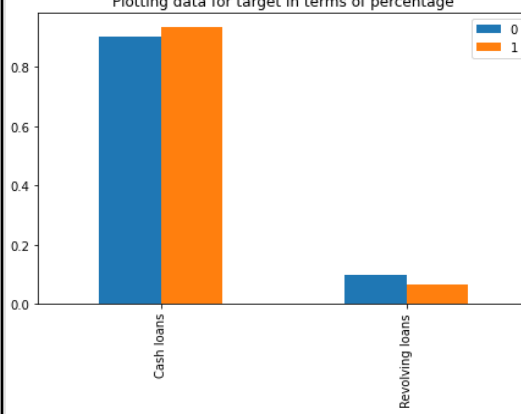
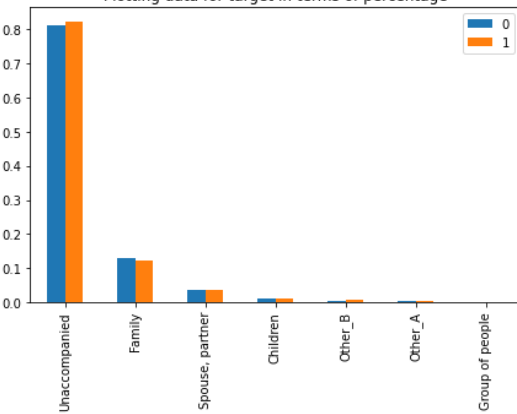
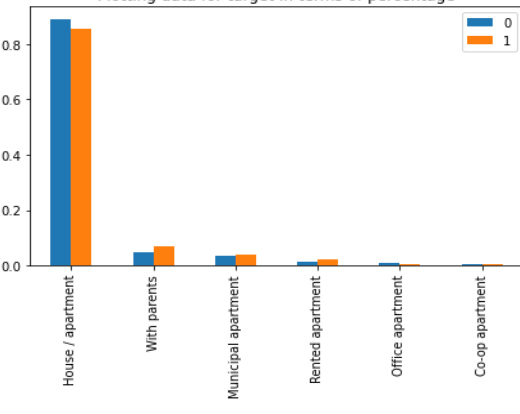
- The target column gives the insight on the clients with payment difficulties (target = 1) and other clients (target = 0).
- On analysis, it can be seen that out of the total 307511 clients, 24825 clients have payment difficulties, i.e. 8.07% clients have payment difficulties.

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.																																																																											
<div>Plotting data for the column: CODE_GENDER</div> <div>Plotting data for target in terms of percentage</div>  <table><tr><th>Gender</th><th>0</th><th>1</th></tr><tr><td>F</td><td>0.68</td><td>0.58</td></tr><tr><td>M</td><td>0.34</td><td>0.44</td></tr></table>	Gender	0	1	F	0.68	0.58	M	0.34	0.44	<div>Plotting data for the column: NAME_EDUCATION_TYPE</div> <div>Plotting data for target in terms of percentage</div>  <table><tr><th>Education Type</th><th>0</th><th>1</th></tr><tr><td>Secondary / secondary special</td><td>0.72</td><td>0.78</td></tr><tr><td>Higher education</td><td>0.25</td><td>0.18</td></tr><tr><td>Incomplete higher</td><td>0.03</td><td>0.04</td></tr><tr><td>Lower secondary</td><td>0.01</td><td>0.02</td></tr><tr><td>Academic degree</td><td>0.00</td><td>0.00</td></tr></table>	Education Type	0	1	Secondary / secondary special	0.72	0.78	Higher education	0.25	0.18	Incomplete higher	0.03	0.04	Lower secondary	0.01	0.02	Academic degree	0.00	0.00	<div>Plotting data for the column: NAME_INCOME_TYPE</div> <div>Plotting data for target in terms of percentage</div>  <table><tr><th>Income Type</th><th>0</th><th>1</th></tr><tr><td>Working</td><td>0.52</td><td>0.62</td></tr><tr><td>Commercial associate</td><td>0.24</td><td>0.22</td></tr><tr><td>Pensioner</td><td>0.19</td><td>0.12</td></tr><tr><td>State servant</td><td>0.08</td><td>0.05</td></tr><tr><td>Student</td><td>0.00</td><td>0.00</td></tr><tr><td>Unemployed</td><td>0.00</td><td>0.00</td></tr><tr><td>Businessman</td><td>0.00</td><td>0.00</td></tr><tr><td>Maternity leave</td><td>0.00</td><td>0.00</td></tr></table>	Income Type	0	1	Working	0.52	0.62	Commercial associate	0.24	0.22	Pensioner	0.19	0.12	State servant	0.08	0.05	Student	0.00	0.00	Unemployed	0.00	0.00	Businessman	0.00	0.00	Maternity leave	0.00	0.00	<div>Plotting data for the column: NAME_FAMILY_STATUS</div> <div>Plotting data for target in terms of percentage</div>  <table><tr><th>Family Status</th><th>0</th><th>1</th></tr><tr><td>Married</td><td>0.65</td><td>0.60</td></tr><tr><td>Single / not married</td><td>0.15</td><td>0.18</td></tr><tr><td>Civil marriage</td><td>0.10</td><td>0.12</td></tr><tr><td>Separated</td><td>0.07</td><td>0.07</td></tr><tr><td>Widow</td><td>0.06</td><td>0.04</td></tr><tr><td>Unknown</td><td>0.00</td><td>0.00</td></tr></table>	Family Status	0	1	Married	0.65	0.60	Single / not married	0.15	0.18	Civil marriage	0.10	0.12	Separated	0.07	0.07	Widow	0.06	0.04	Unknown	0.00	0.00
Gender	0	1																																																																												
F	0.68	0.58																																																																												
M	0.34	0.44																																																																												
Education Type	0	1																																																																												
Secondary / secondary special	0.72	0.78																																																																												
Higher education	0.25	0.18																																																																												
Incomplete higher	0.03	0.04																																																																												
Lower secondary	0.01	0.02																																																																												
Academic degree	0.00	0.00																																																																												
Income Type	0	1																																																																												
Working	0.52	0.62																																																																												
Commercial associate	0.24	0.22																																																																												
Pensioner	0.19	0.12																																																																												
State servant	0.08	0.05																																																																												
Student	0.00	0.00																																																																												
Unemployed	0.00	0.00																																																																												
Businessman	0.00	0.00																																																																												
Maternity leave	0.00	0.00																																																																												
Family Status	0	1																																																																												
Married	0.65	0.60																																																																												
Single / not married	0.15	0.18																																																																												
Civil marriage	0.10	0.12																																																																												
Separated	0.07	0.07																																																																												
Widow	0.06	0.04																																																																												
Unknown	0.00	0.00																																																																												

Univariate Analysis of Target

Observations																																																																																																																		
Clients with low income are more probable to take loans and become defaulters.	Clients more likely opt for cash loans than revolving loans while the defaulters are also greater in case of cash loans.	Major loan clients who were unaccompanied, took loans and later had difficulties in payment.	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.																																																																																																															
<p>Plotting data for the column: OCCUPATION_TYPE</p> <p>Plotting data for target in terms of percentage</p>  <table><caption>Approximate data for OCCUPATION_TYPE</caption><thead><tr><th>Occupation Type</th><th>Target 0 (%)</th><th>Target 1 (%)</th></tr></thead><tbody><tr><td>Laborers</td><td>0.48</td><td>0.49</td></tr><tr><td>Sales staff</td><td>0.10</td><td>0.12</td></tr><tr><td>Core staff</td><td>0.08</td><td>0.06</td></tr><tr><td>Managers</td><td>0.06</td><td>0.05</td></tr><tr><td>Drivers</td><td>0.05</td><td>0.08</td></tr><tr><td>High skill tech staff</td><td>0.03</td><td>0.02</td></tr><tr><td>Accountants</td><td>0.02</td><td>0.01</td></tr><tr><td>Medicine staff</td><td>0.02</td><td>0.01</td></tr><tr><td>Security staff</td><td>0.02</td><td>0.01</td></tr><tr><td>Cooking staff</td><td>0.02</td><td>0.01</td></tr><tr><td>Cleaning staff</td><td>0.01</td><td>0.01</td></tr><tr><td>Private service staff</td><td>0.01</td><td>0.01</td></tr><tr><td>Low-skill Laborers</td><td>0.01</td><td>0.01</td></tr><tr><td>Secretaries</td><td>0.01</td><td>0.01</td></tr><tr><td>Walters/barmen staff</td><td>0.01</td><td>0.01</td></tr><tr><td>Reality agents</td><td>0.01</td><td>0.01</td></tr><tr><td>HR staff</td><td>0.01</td><td>0.01</td></tr><tr><td>IT staff</td><td>0.01</td><td>0.01</td></tr></tbody></table>	Occupation Type	Target 0 (%)	Target 1 (%)	Laborers	0.48	0.49	Sales staff	0.10	0.12	Core staff	0.08	0.06	Managers	0.06	0.05	Drivers	0.05	0.08	High skill tech staff	0.03	0.02	Accountants	0.02	0.01	Medicine staff	0.02	0.01	Security staff	0.02	0.01	Cooking staff	0.02	0.01	Cleaning staff	0.01	0.01	Private service staff	0.01	0.01	Low-skill Laborers	0.01	0.01	Secretaries	0.01	0.01	Walters/barmen staff	0.01	0.01	Reality agents	0.01	0.01	HR staff	0.01	0.01	IT staff	0.01	0.01	<p>Plotting data for the column: NAME_CONTRACT_TYPE</p> <p>Plotting data for target in terms of percentage</p>  <table><caption>Approximate data for NAME_CONTRACT_TYPE</caption><thead><tr><th>Contract Type</th><th>Target 0 (%)</th><th>Target 1 (%)</th></tr></thead><tbody><tr><td>Cash loans</td><td>0.88</td><td>0.92</td></tr><tr><td>Revolving loans</td><td>0.10</td><td>0.08</td></tr></tbody></table>	Contract Type	Target 0 (%)	Target 1 (%)	Cash loans	0.88	0.92	Revolving loans	0.10	0.08	<p>Plotting data for the column: NAME_TYPE_SUITE</p> <p>Plotting data for target in terms of percentage</p>  <table><caption>Approximate data for NAME_TYPE_SUITE</caption><thead><tr><th>Suite Type</th><th>Target 0 (%)</th><th>Target 1 (%)</th></tr></thead><tbody><tr><td>Unaccompanied</td><td>0.80</td><td>0.82</td></tr><tr><td>Family</td><td>0.12</td><td>0.11</td></tr><tr><td>Spouse, partner</td><td>0.03</td><td>0.02</td></tr><tr><td>Children</td><td>0.01</td><td>0.01</td></tr><tr><td>Other_B</td><td>0.01</td><td>0.01</td></tr><tr><td>Other_A</td><td>0.01</td><td>0.01</td></tr><tr><td>Group of people</td><td>0.01</td><td>0.01</td></tr></tbody></table>	Suite Type	Target 0 (%)	Target 1 (%)	Unaccompanied	0.80	0.82	Family	0.12	0.11	Spouse, partner	0.03	0.02	Children	0.01	0.01	Other_B	0.01	0.01	Other_A	0.01	0.01	Group of people	0.01	0.01	<p>Plotting data for the column: NAME_HOUSING_TYPE</p> <p>Plotting data for target in terms of percentage</p>  <table><caption>Approximate data for NAME_HOUSING_TYPE</caption><thead><tr><th>Housing Type</th><th>Target 0 (%)</th><th>Target 1 (%)</th></tr></thead><tbody><tr><td>House / apartment</td><td>0.88</td><td>0.85</td></tr><tr><td>With parents</td><td>0.05</td><td>0.08</td></tr><tr><td>Municipal apartment</td><td>0.03</td><td>0.04</td></tr><tr><td>Rented apartment</td><td>0.02</td><td>0.02</td></tr><tr><td>Office apartment</td><td>0.01</td><td>0.01</td></tr><tr><td>Co-op apartment</td><td>0.01</td><td>0.01</td></tr></tbody></table>	Housing Type	Target 0 (%)	Target 1 (%)	House / apartment	0.88	0.85	With parents	0.05	0.08	Municipal apartment	0.03	0.04	Rented apartment	0.02	0.02	Office apartment	0.01	0.01	Co-op apartment	0.01	0.01
Occupation Type	Target 0 (%)	Target 1 (%)																																																																																																																
Laborers	0.48	0.49																																																																																																																
Sales staff	0.10	0.12																																																																																																																
Core staff	0.08	0.06																																																																																																																
Managers	0.06	0.05																																																																																																																
Drivers	0.05	0.08																																																																																																																
High skill tech staff	0.03	0.02																																																																																																																
Accountants	0.02	0.01																																																																																																																
Medicine staff	0.02	0.01																																																																																																																
Security staff	0.02	0.01																																																																																																																
Cooking staff	0.02	0.01																																																																																																																
Cleaning staff	0.01	0.01																																																																																																																
Private service staff	0.01	0.01																																																																																																																
Low-skill Laborers	0.01	0.01																																																																																																																
Secretaries	0.01	0.01																																																																																																																
Walters/barmen staff	0.01	0.01																																																																																																																
Reality agents	0.01	0.01																																																																																																																
HR staff	0.01	0.01																																																																																																																
IT staff	0.01	0.01																																																																																																																
Contract Type	Target 0 (%)	Target 1 (%)																																																																																																																
Cash loans	0.88	0.92																																																																																																																
Revolving loans	0.10	0.08																																																																																																																
Suite Type	Target 0 (%)	Target 1 (%)																																																																																																																
Unaccompanied	0.80	0.82																																																																																																																
Family	0.12	0.11																																																																																																																
Spouse, partner	0.03	0.02																																																																																																																
Children	0.01	0.01																																																																																																																
Other_B	0.01	0.01																																																																																																																
Other_A	0.01	0.01																																																																																																																
Group of people	0.01	0.01																																																																																																																
Housing Type	Target 0 (%)	Target 1 (%)																																																																																																																
House / apartment	0.88	0.85																																																																																																																
With parents	0.05	0.08																																																																																																																
Municipal apartment	0.03	0.04																																																																																																																
Rented apartment	0.02	0.02																																																																																																																
Office apartment	0.01	0.01																																																																																																																
Co-op apartment	0.01	0.01																																																																																																																

Univariate Analysis of Target

Observations		
<p>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.</p>	<p>Clients who do not own cars are more likely to take loans. Notably, the percentage of defaulters is also higher in these cases.</p>	<p>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.</p>
<p>Plotting data for the column: ORGANIZATION_TYPE</p> <p>Plotting data for target in terms of percentage</p>	<p>Plotting data for the column: FLAG_OWN_CAR</p> <p>Plotting data for target in terms of percentage</p>	<p>Plotting data for the column: FLAG_OWN_REALTY</p> <p>Plotting data for target in terms of percentage</p>

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

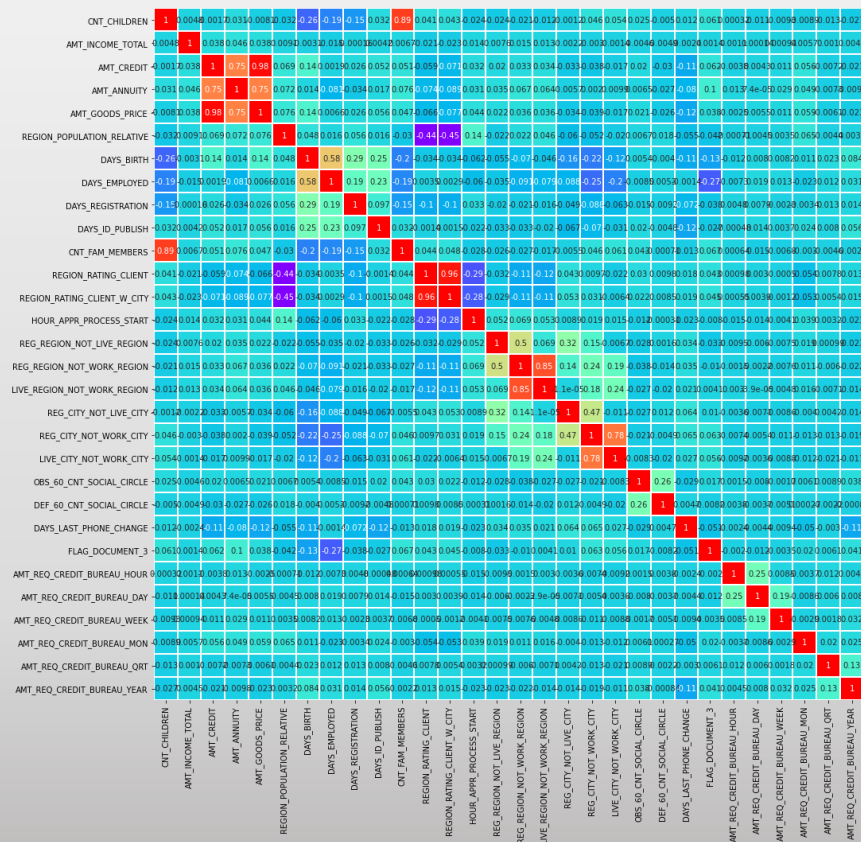
Clients with payment difficulties

	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

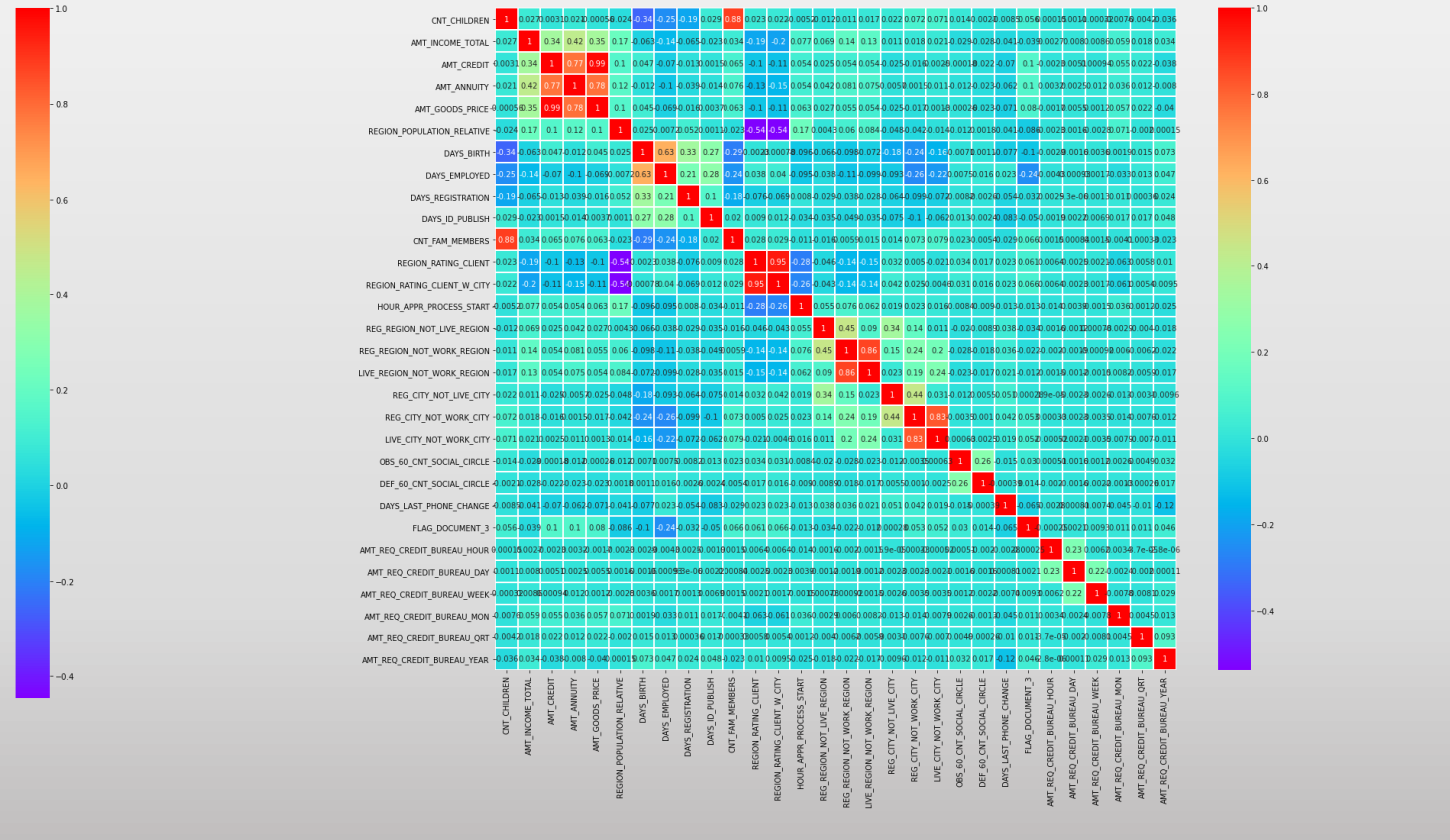
Clients without payment difficulties

Bivariate Analysis of Target

Heatmaps showing correlation between different variables



Clients with payment difficulties



Clients without payment difficulties

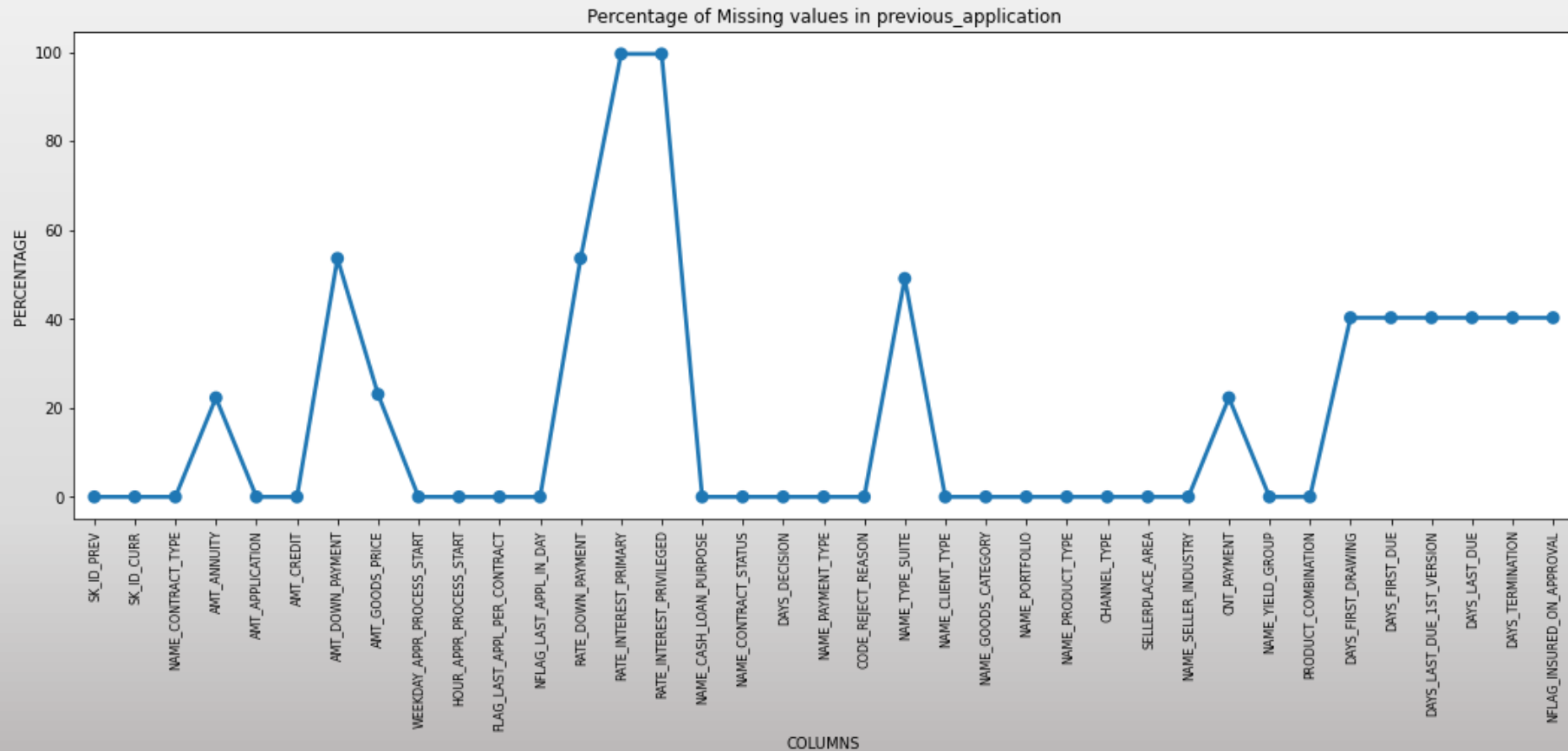
Bivariate Analysis of Target

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%).

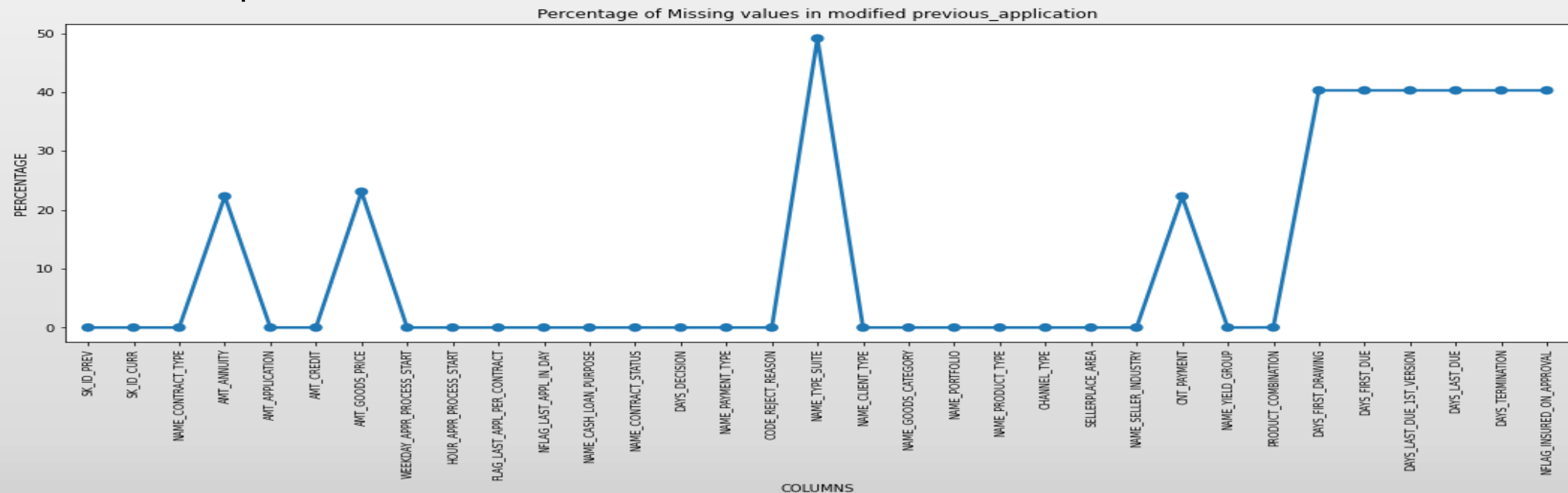
Finding Missing Values in Previous Application Dataset

- 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

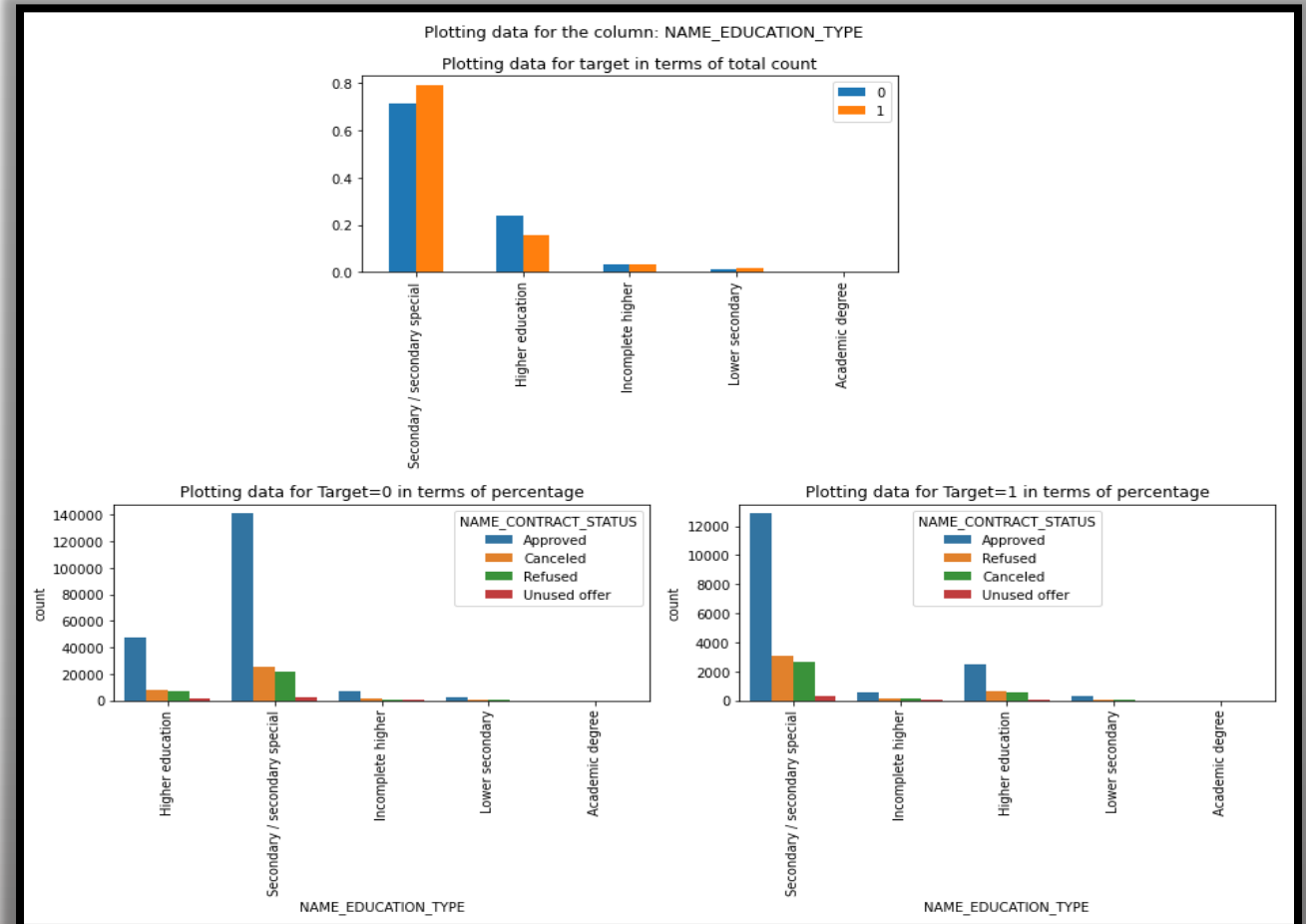
- Next we have to impute the missing values in the columns that have lower missing value counts.
- Columns that need missing value treatment :

COLUMN INDEX
AMT_ANNUIITY
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.

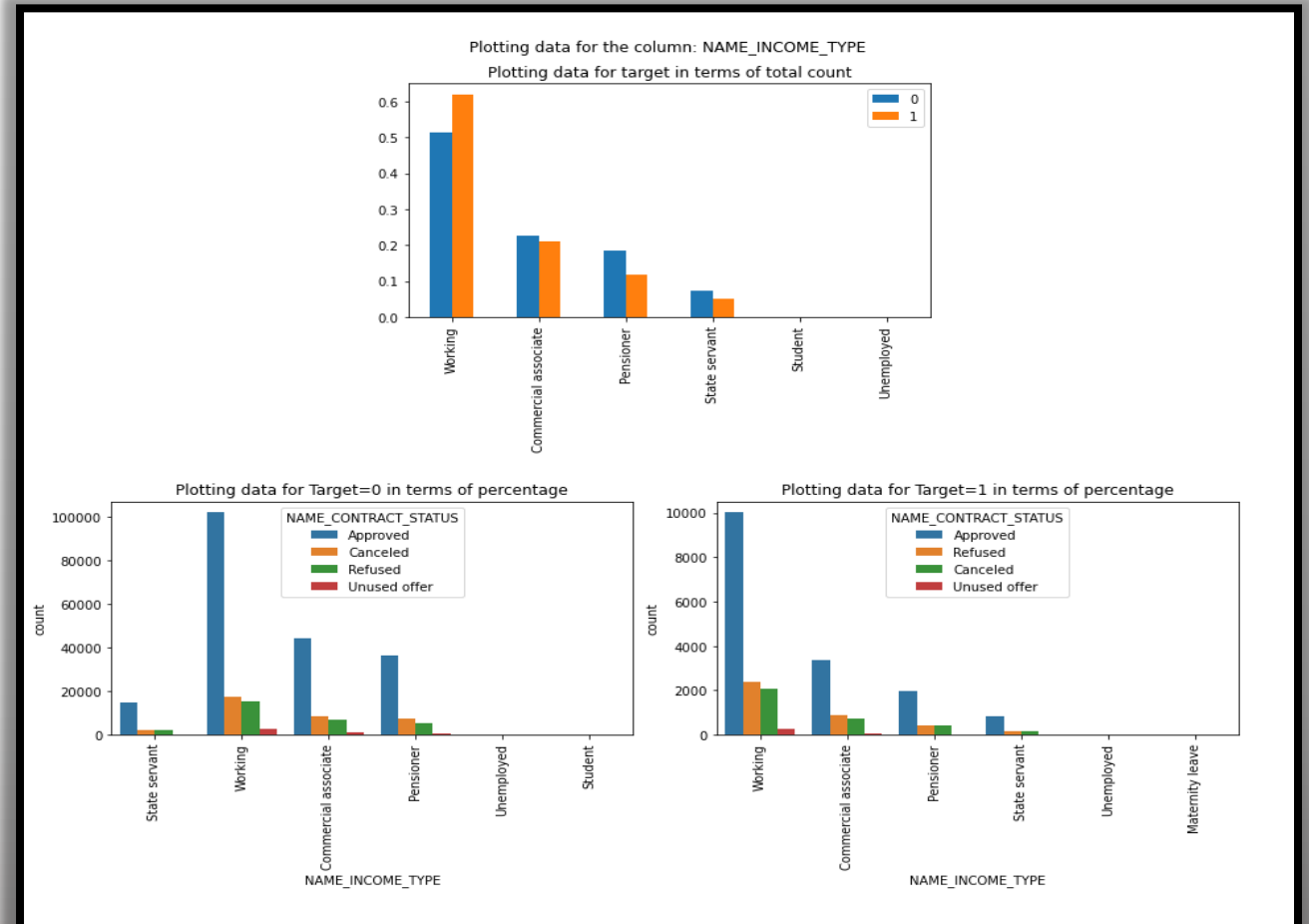
Univariate Analysis of Target in Merged Dataset

- Clients with secondary and higher education mostly opt for loans.
- Maximum number of secondary educated and higher educated clients with approved loans are defaulters.



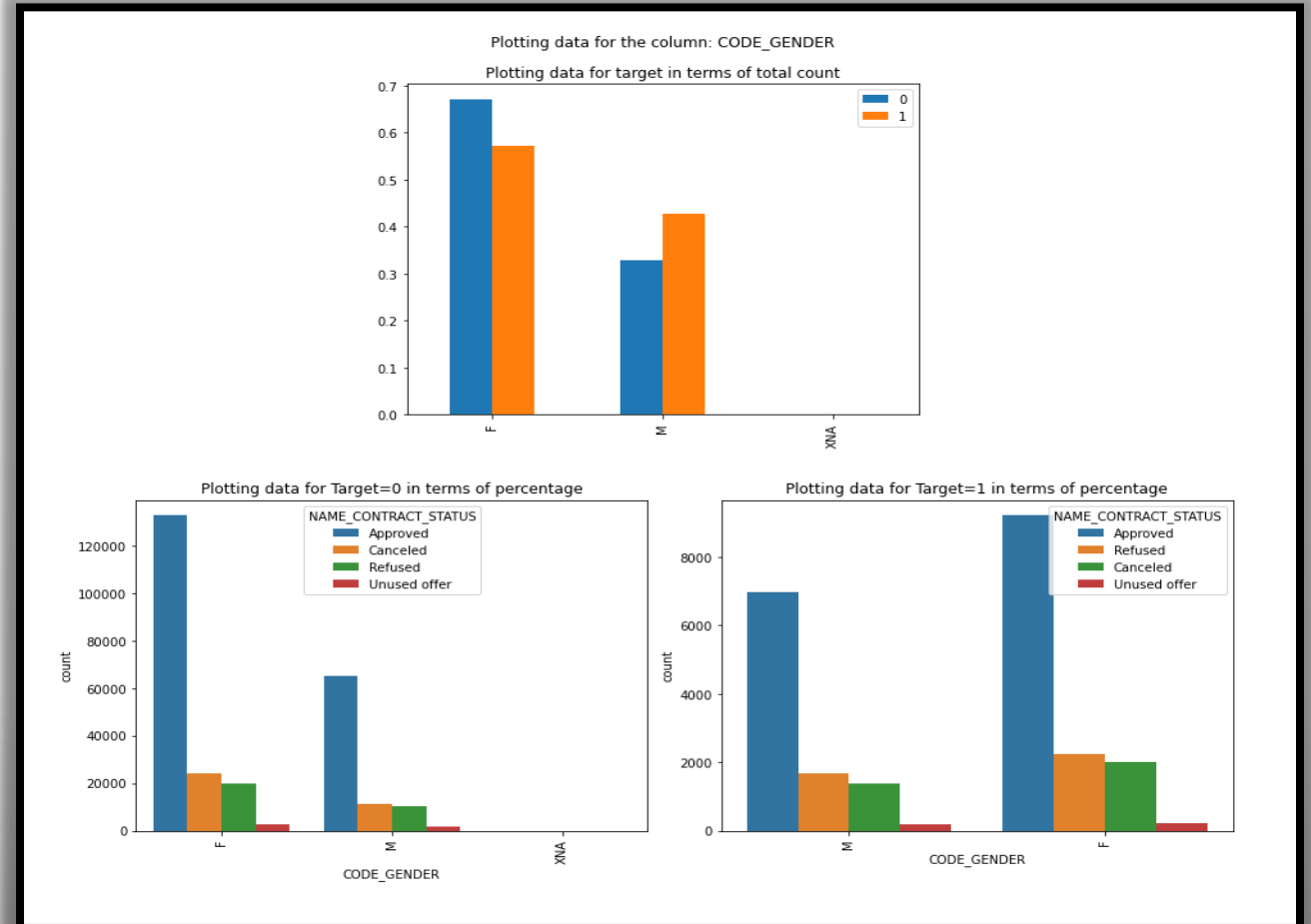
Univariate Analysis of Target in Merged Dataset

- Working professionals, pensioners and commercial associates mostly get loans approved.
- Pensioners are less likely to be defaulters as compared to working professionals.



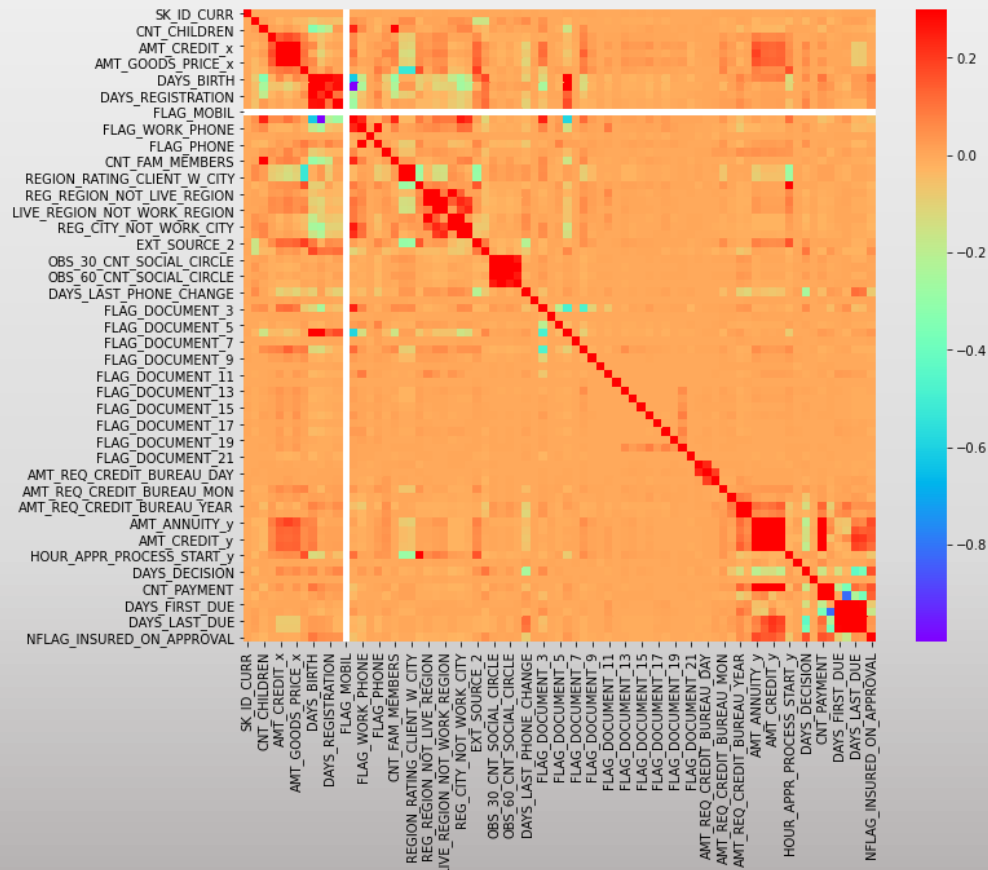
Univariate Analysis of Target in Merged Dataset

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

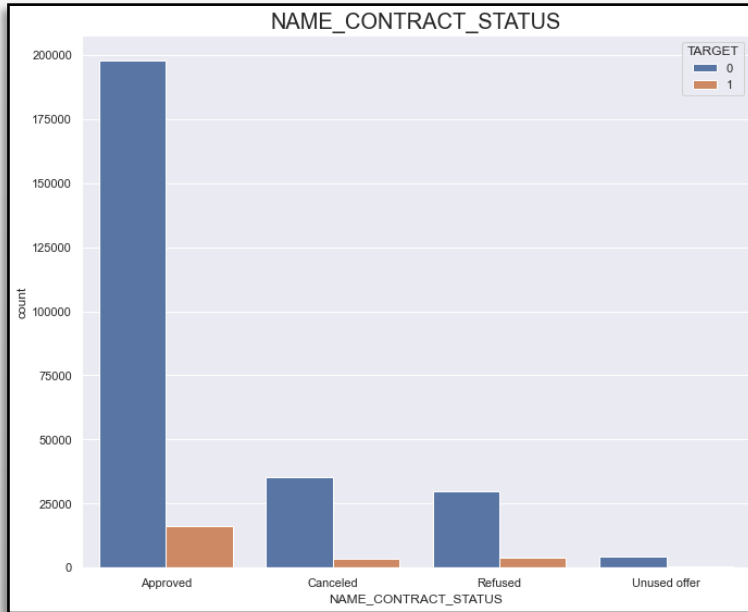


Bivariate Analysis of Merged Dataset

- 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



		Number	Percentage (%)
NAME_CONTRACT_STATUS	TARGET		
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.