Credit EDA Assignment

Ihsan Qaid Jowhar

Objective

- When the company receives a loan application, the company must decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
  - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

- This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected.

Approach

- The application dataset and previous application datasets are cleaned by:
  - Identifying columns with highest null percentages and dropping those columns, in some cases the rows are dropped or imputed with mean/median
  - Outliers in some columns are not removed, rather there are replaced with the upper / lower fence values.
  - Incorrect values like negative numbers in some columns are converted to absolute values.

- Analysis (univariate, bivariate and multivariate) is performed separately on both datasets

- Both datasets are merged to a single 'Merged' dataframe.

- Analysis (univariate, bivariate and multivariate) is performed on the 'Merged' dataframe
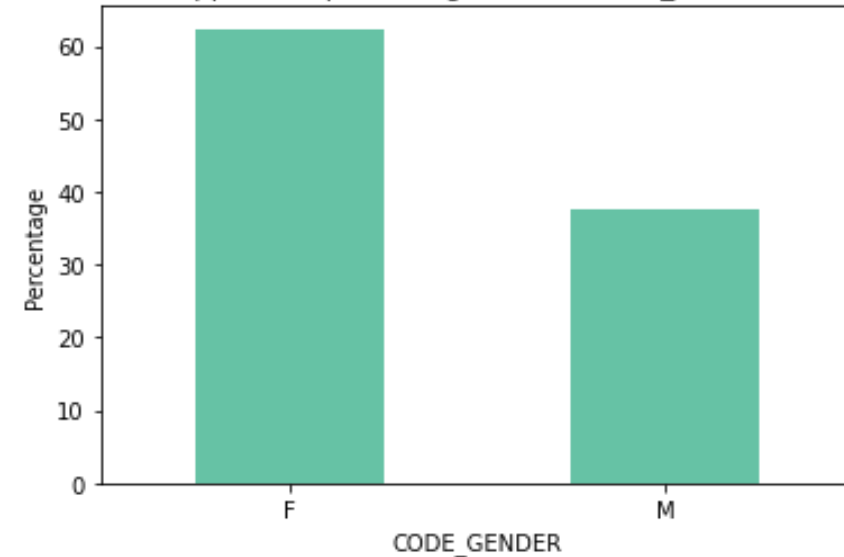
APPLICATIONS DATASET

UNIVARIATE ANALYSIS
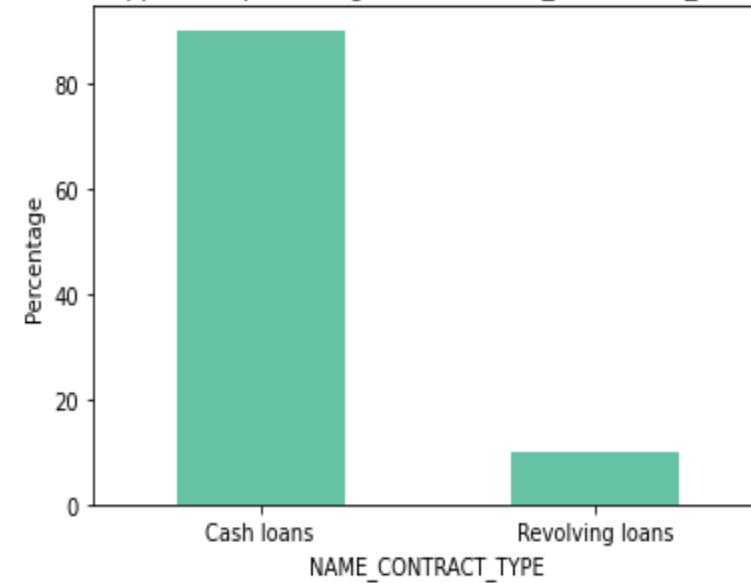

Applicants percentage across TARGET
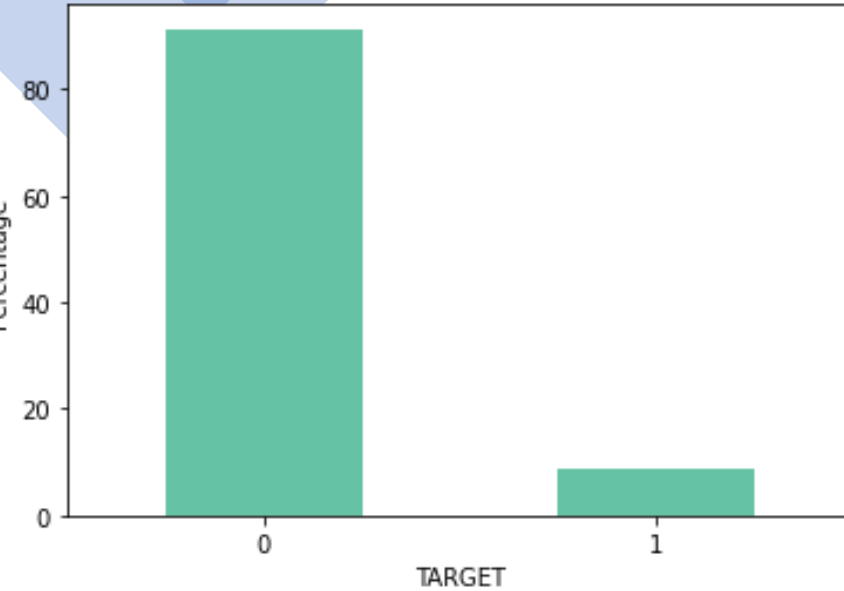
More than **90%** of applicants are non-defaulters

Applicants percentage across NAME_CONTRACT_TYPE
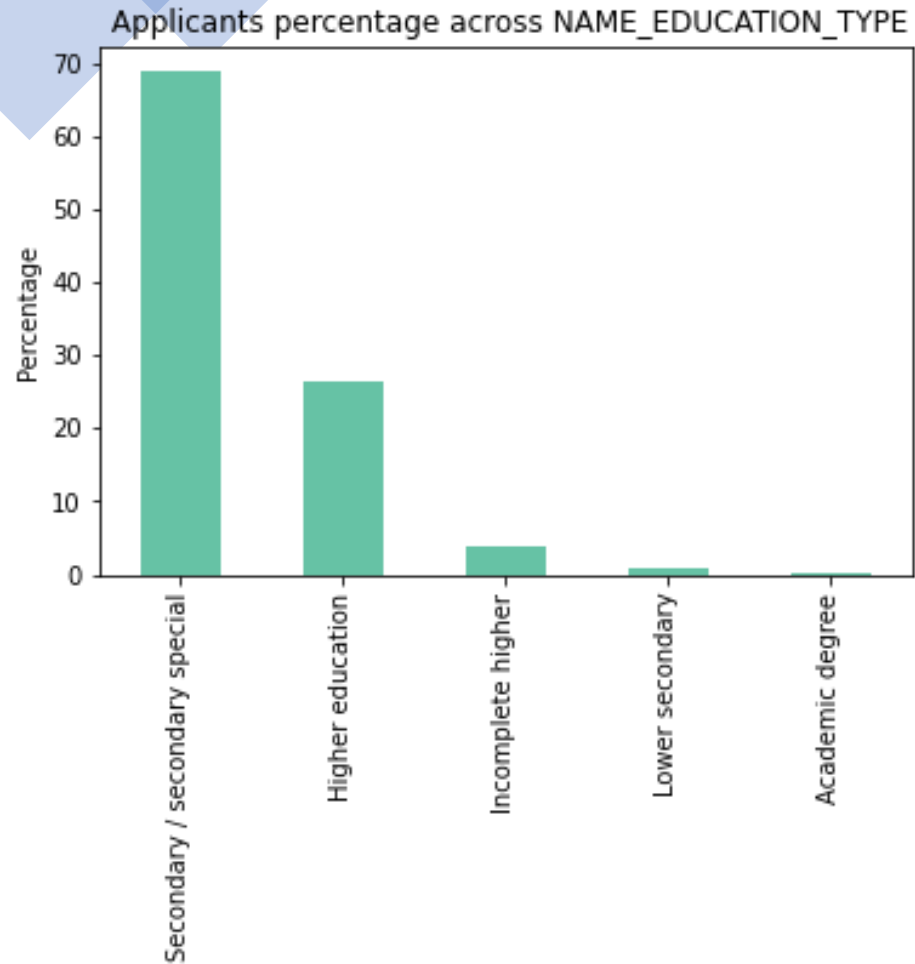
More than **80%** of loans are cash loans

Applicants percentage across CODE_GENDER

Females form the majority of loan applicants **(>60%)**

UNIVARIATE ANALYSIS

Applicants percentage across NAME_EDUCATION_TYPE
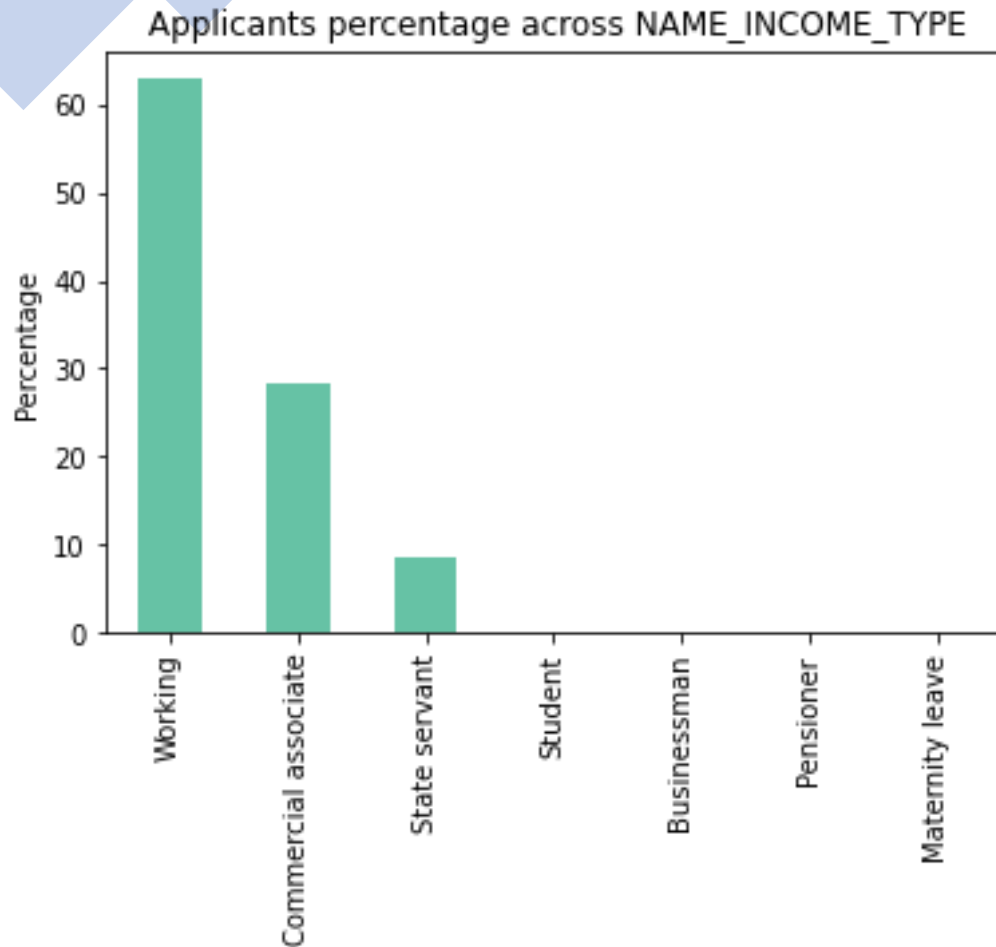

Applicants percentage across NAME_FAMILY_STATUS

Almost 70% of applicants have secondary level education
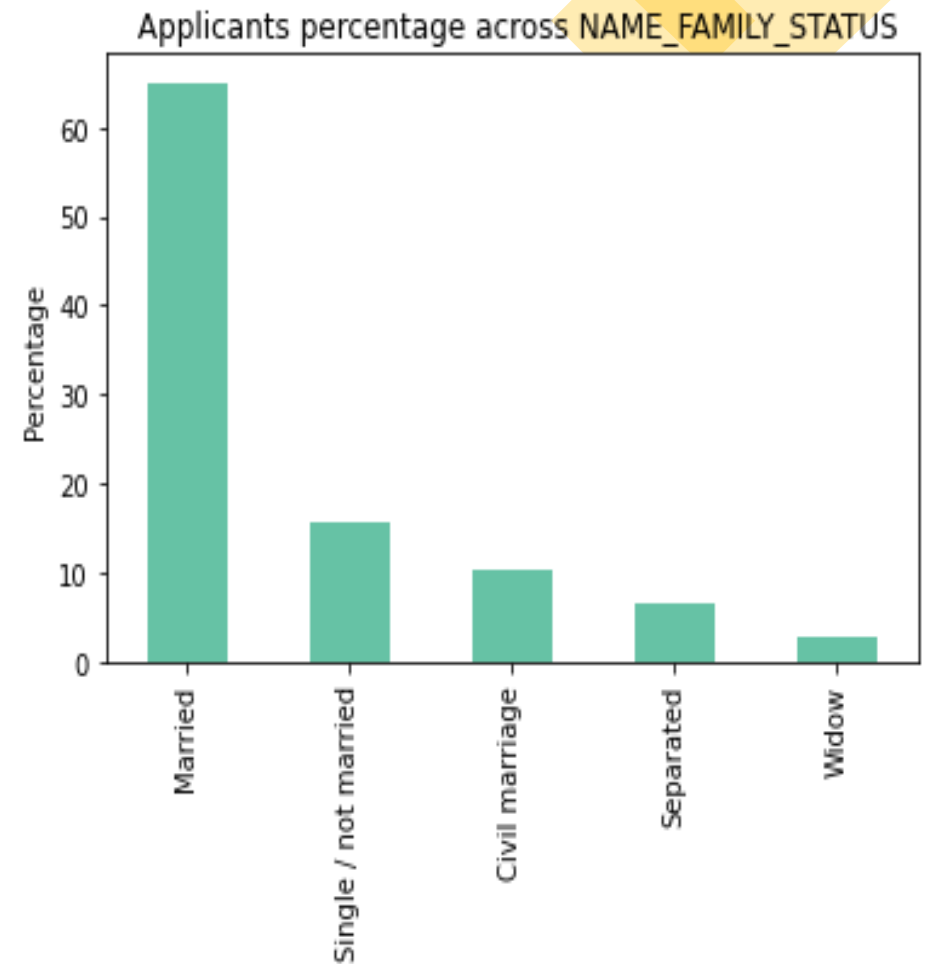
More than 60% of applicants are married whereas only about 18% are single

UNIVARIATE ANALYSIS

Applicants percentage across NAME_INCOME_TYPE

Applicants percentage across OCCUPATION_TYPE

Majority of the applicants (>60%) are working, with negligible numbers in Student and Businessmen
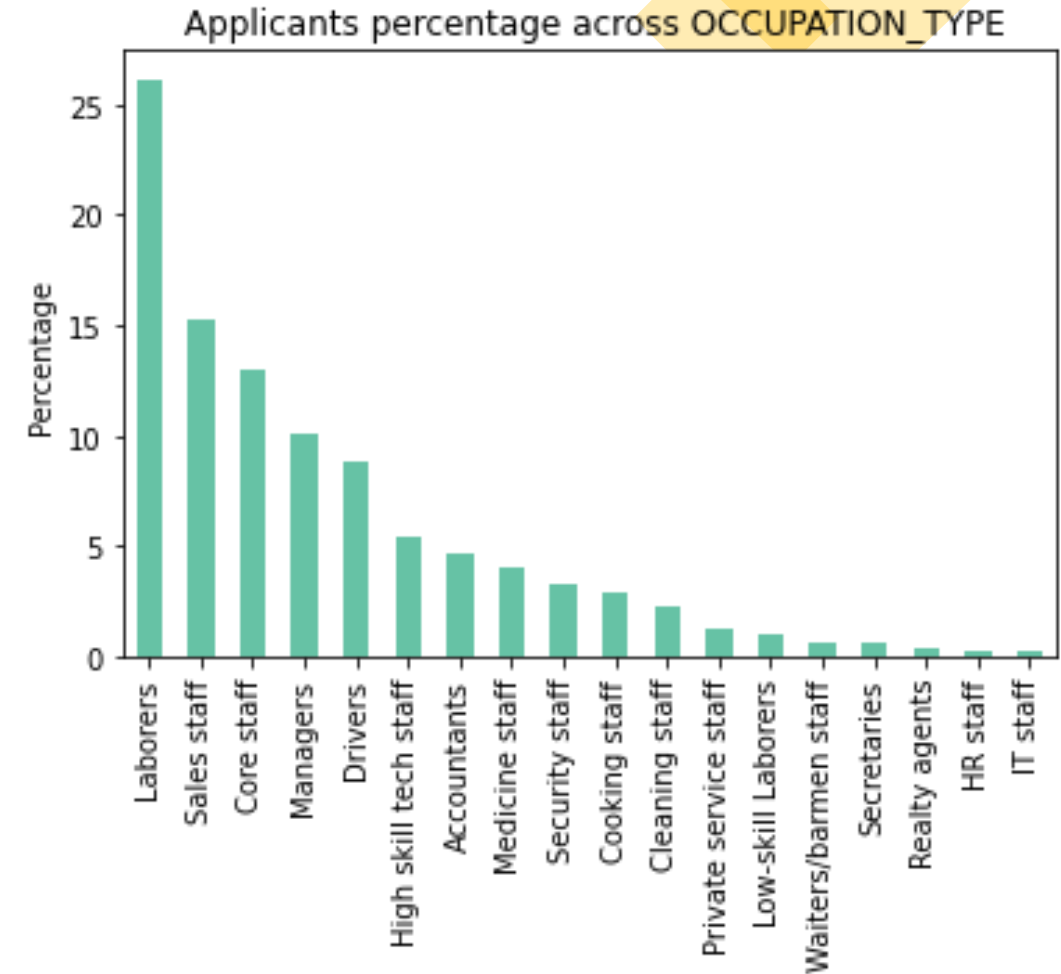
25% of applicants are laborers

BIVARIATE ANALYSIS
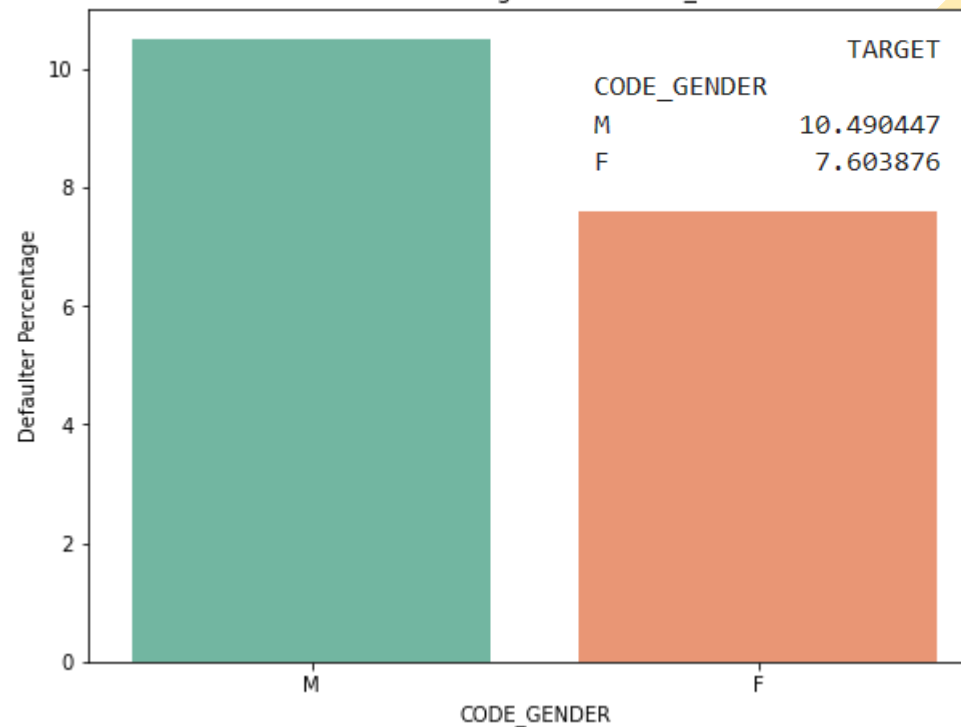
Even though more than 60% of the applicants are Females (as seen in the univariate analysis), the defaulter percentage in Males is very close to Females (7.6% vs 10.5%)

Recommendation : Going by the proportion of defaulters to applicants, it's safer to give a loan to Females

BIVARIATE ANALYSIS

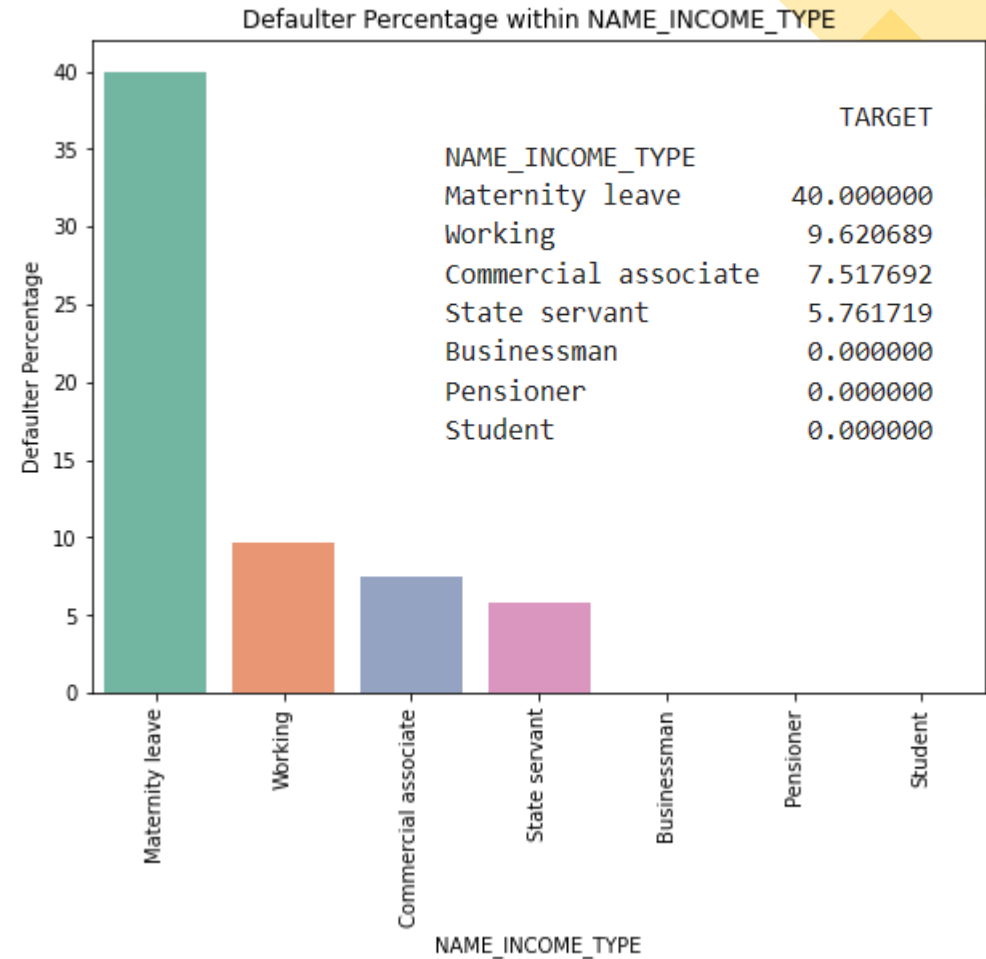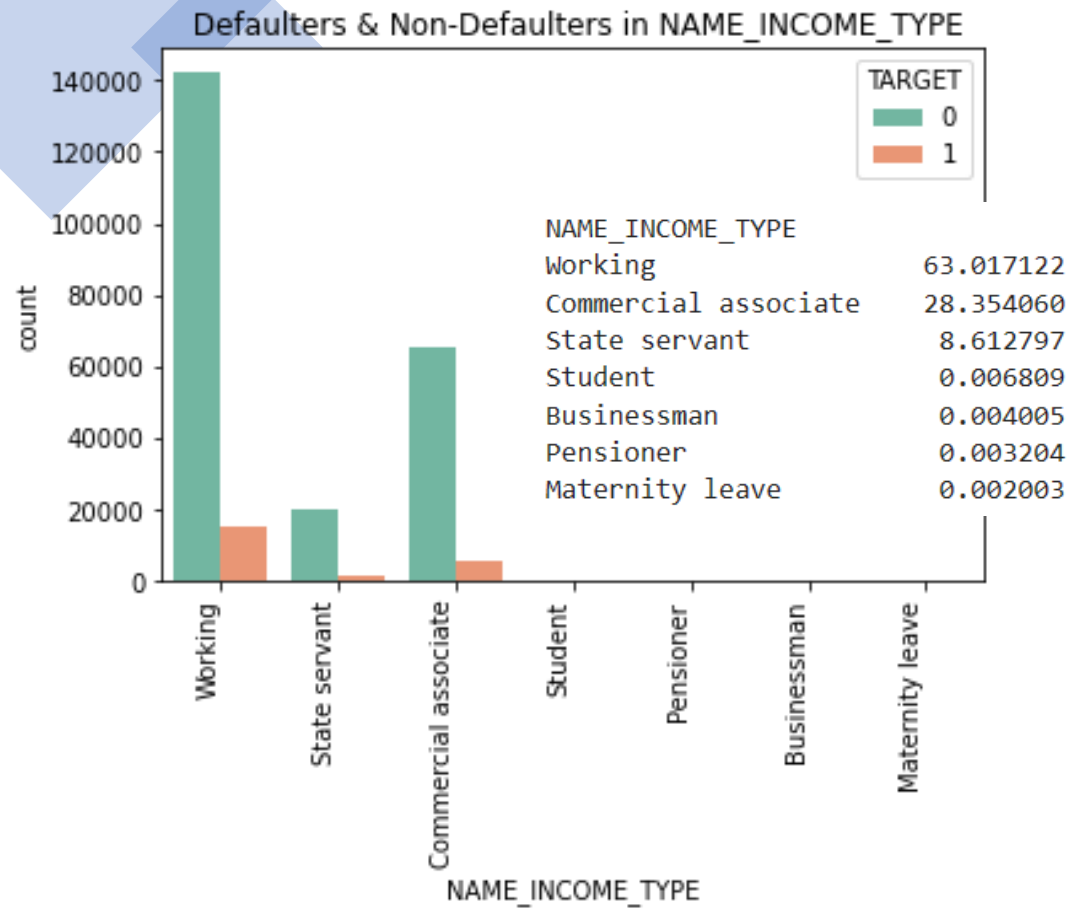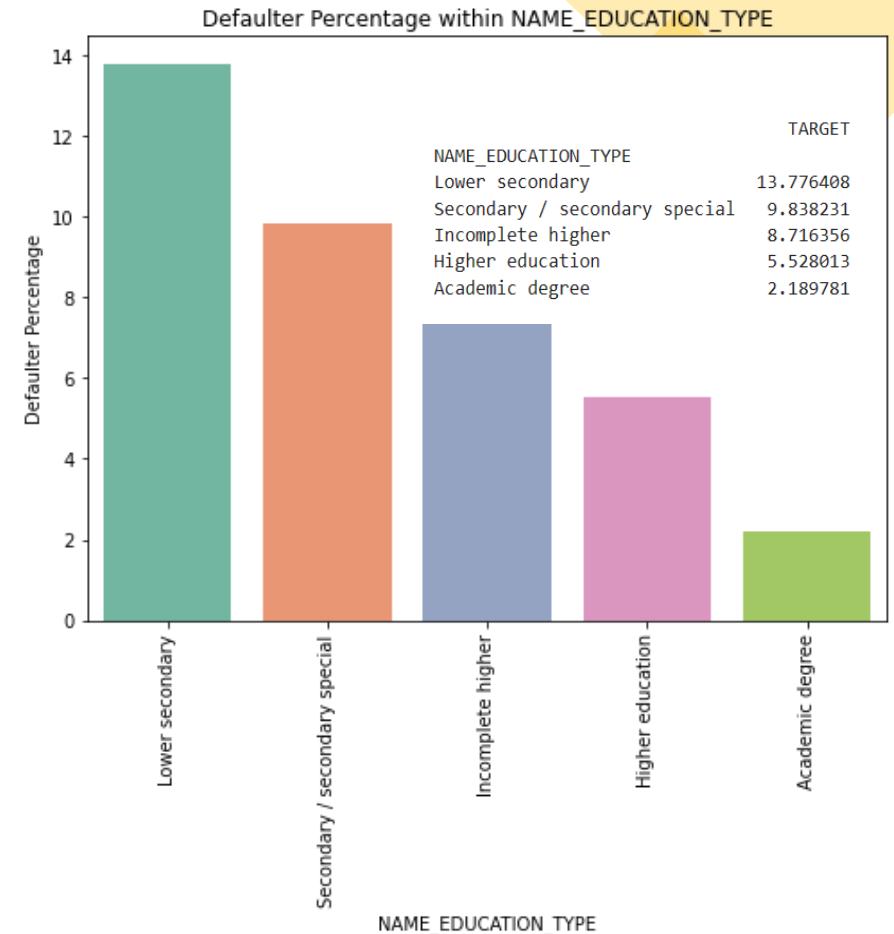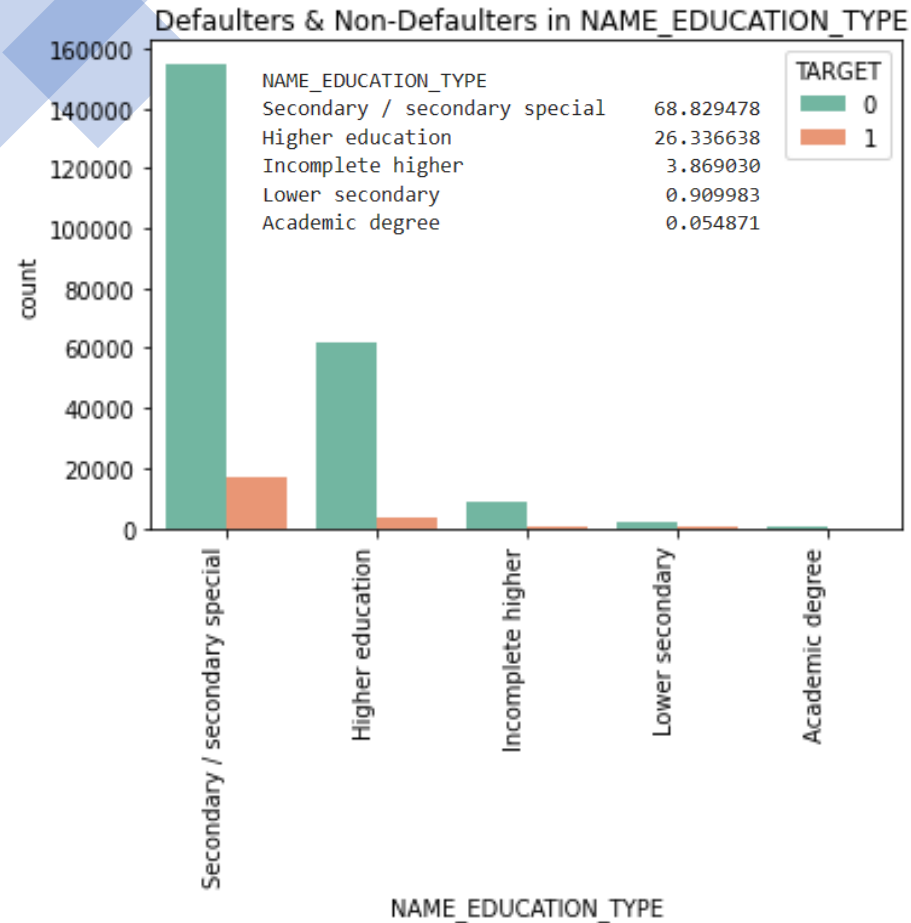Defaulters & Non-Defaulters in NAME_INCOME_TYPE

NAME_INCOME_TYPE	
Working	63.017122
Commercial associate	28.354060
State servant	8.612797
Student	0.006809
Businessman	0.004005
Pensioner	0.003204
Maternity leave	0.002003

Defaulter Percentage within NAME_INCOME_TYPE

NAME_INCOME_TYPE	TARGET
Maternity leave	40.000000
Working	9.620689
Commercial associate	7.517692
State servant	5.761719
Businessman	0.000000
Pensioner	0.000000
Student	0.000000

More than 60% of the applicants are in the Working category of income type (as seen in the univariate analysis), but the defaulter percentage within applicants on Maternity Leave is a clear majority (40%)

Recommendation : Going by the proportion of defaulters to applicants, it's safer to give a loan to Working people as they take up more than 60% of applications, yet their defaulter percentage is low. Also, Businessman, Pensioners and Students have 0% default rate
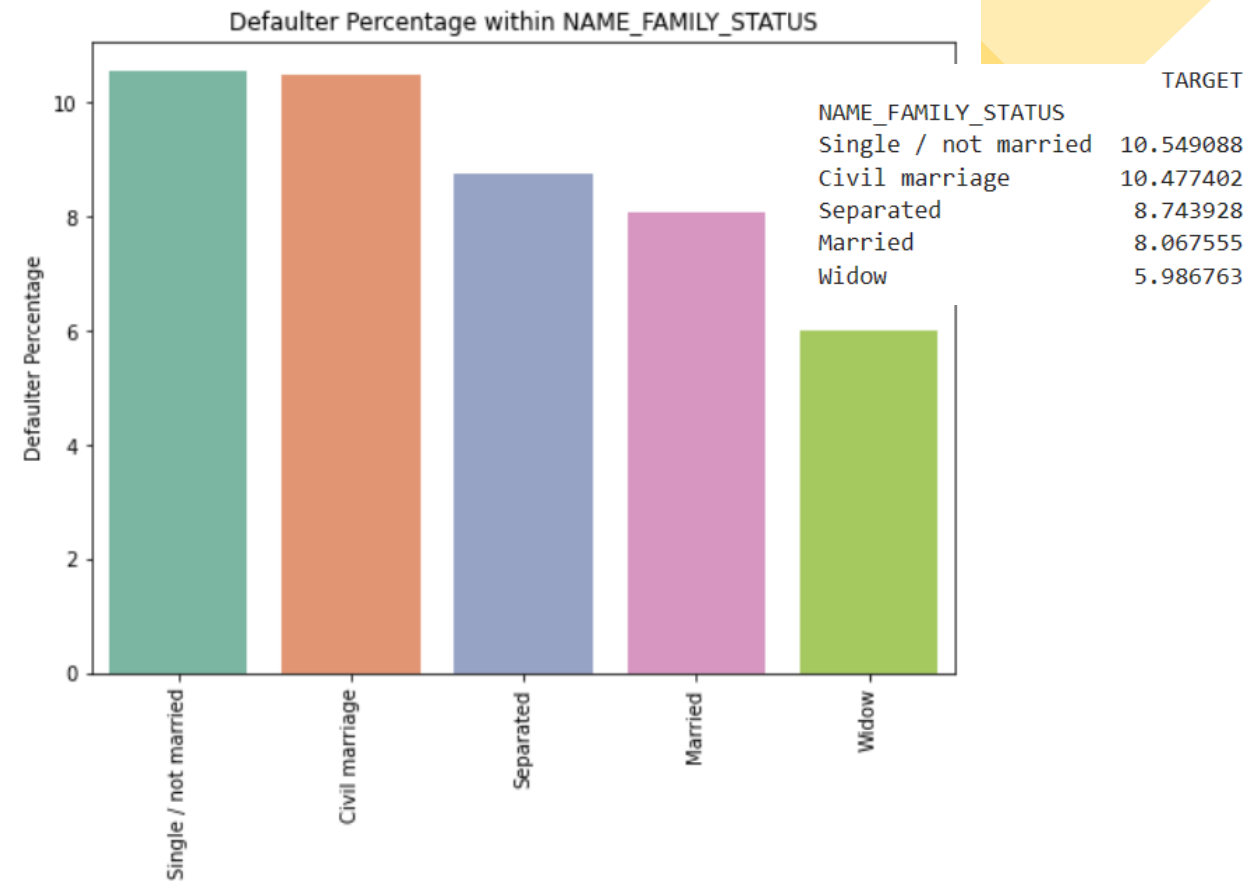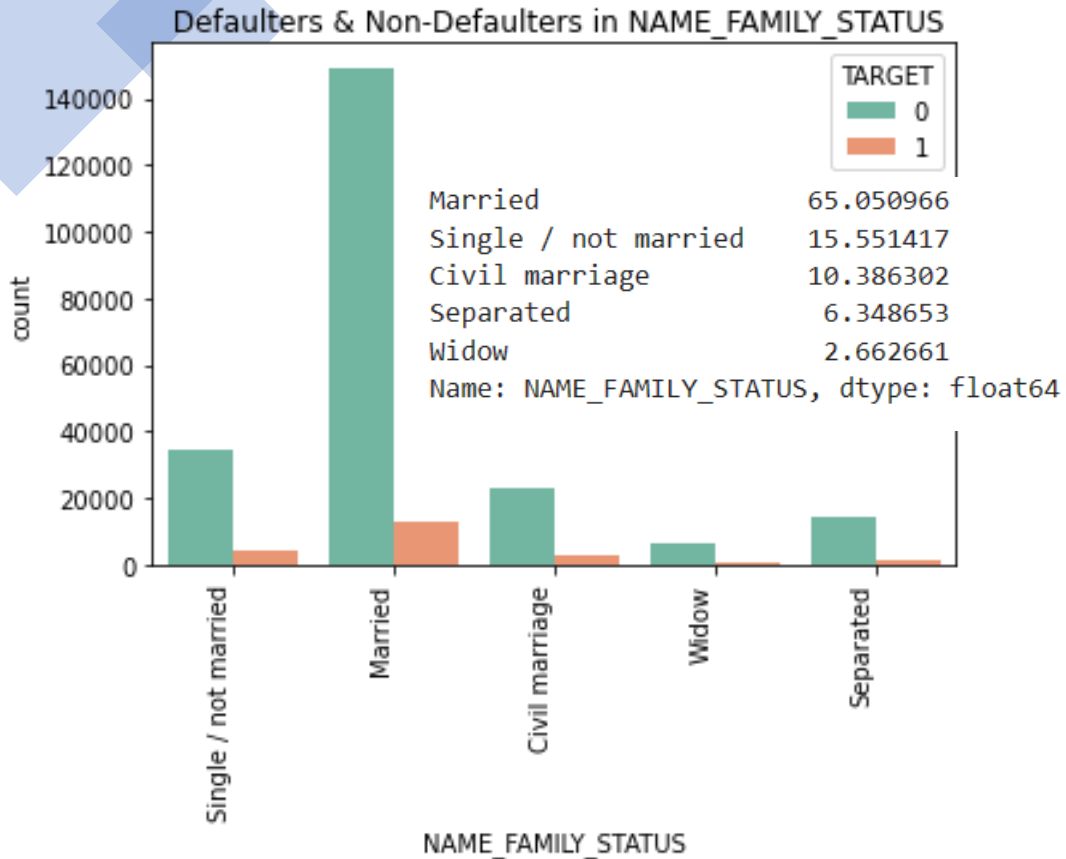
BIVARIATE ANALYSIS

Defaulters & Non-Defaulters in NAME_EDUCATION_TYPE

NAME_EDUCATION_TYPE	
Secondary / secondary special	68.829478
Higher education	26.336638
Incomplete higher	3.869030
Lower secondary	0.909983
Academic degree	0.054871

TARGET: 0, 1

Defaulter Percentage within NAME_EDUCATION_TYPE

NAME_EDUCATION_TYPE	TARGET
Lower secondary	13.776408
Secondary / secondary special	9.838231
Incomplete higher	8.716356
Higher education	5.528013
Academic degree	2.189781

Almost 70% of the applicants have Secondary level education (as seen in the univariate analysis) as compared to Lower secondary which is less than 1%, but applicants within Lower Secondary category have a defaulting percent of 13.7%

Recommendation : Going by the proportion of defaulters to applicants, it's safer to give a loan to applicants with Secondary education or Higher education as they take up more than 95% of applications (combined), yet their defaulter percentage is only about 15% combined as compared to 13.7% amongst Lower Secondary

BIVARIATE ANALYSIS

Defaulters & Non-Defaulters in NAME_FAMILY_STATUS

TARGET	
0	
1	

Married	65.050966
Single / not married	15.551417
Civil marriage	10.386302
Separated	6.348653
Widow	2.662661

Name: NAME_FAMILY_STATUS, dtype: float64

Defaulter Percentage within NAME_FAMILY_STATUS

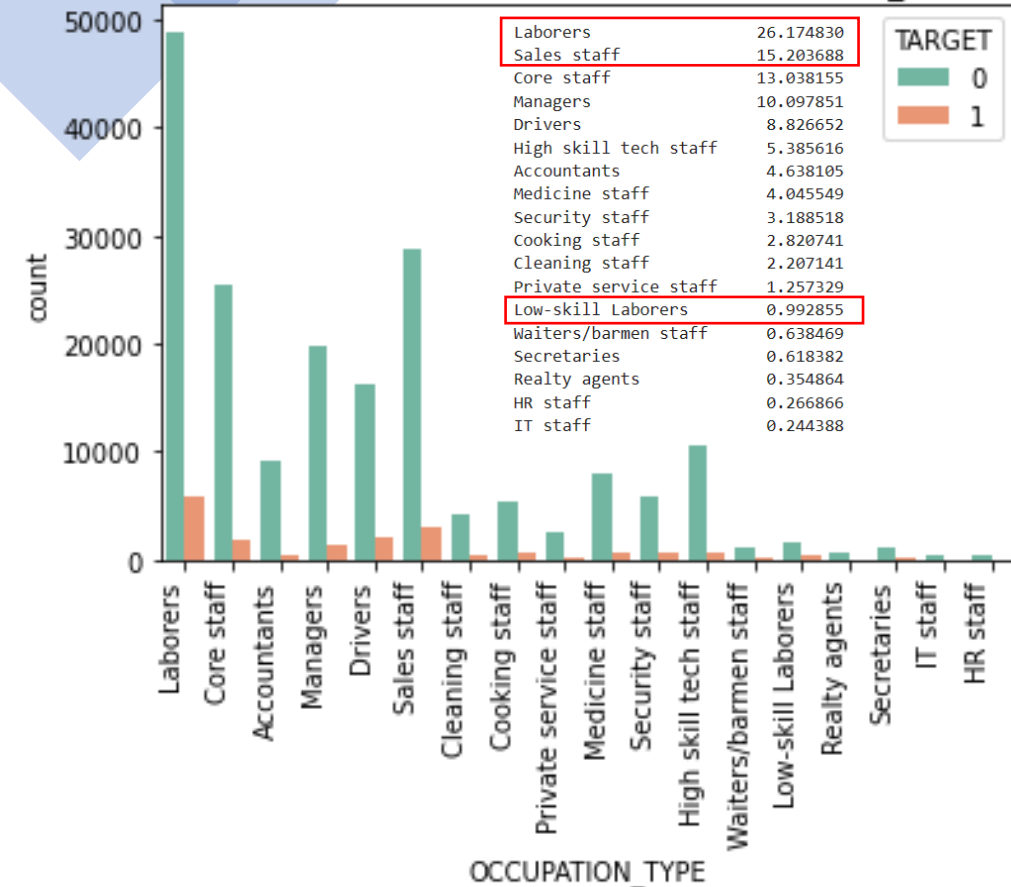|	TARGET
NAME_FAMILY_STATUS	
Single / not married	10.549088
Civil marriage	10.477402
Separated	8.743928
Married	8.067555
Widow	5.986763

More than 65% of the applicants are Married (as seen in the univariate analysis) as compared to Single people who are 15%, but applicants who are Single, have a defaulting percent of 10.5%
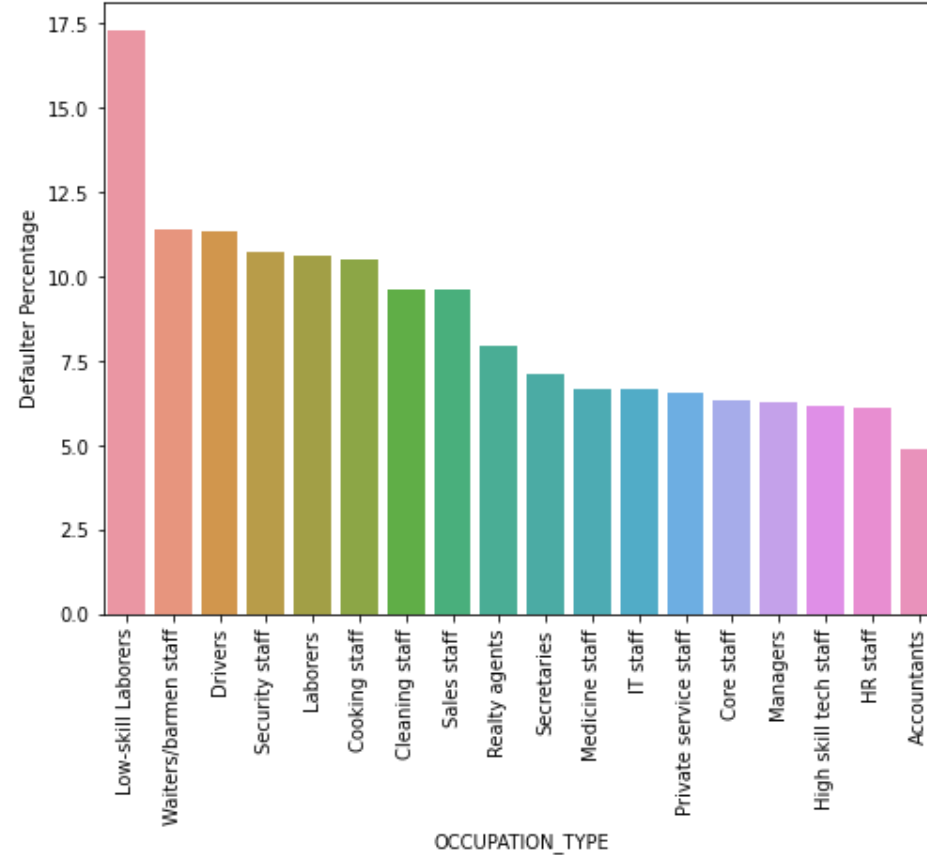
Recommendation : Going by the proportion of defaulters to applicants, it's safer to give a loan to applicants who are Married as they take up more than 65% yet their defaulter percentage is only about 8% as compared to 10.5% amongst Single people who only take up 15% of applications.

BIVARIATE ANALYSIS

Defaulters & Non-Defaulters in OCCUPATION_TYPE

OCCUPATION_TYPE	Value
Laborers	26.174830
Sales staff	15.203688
Core staff	13.038155
Managers	10.097851
Drivers	8.826652
High skill tech staff	5.385616
Accountants	4.638105
Medicine staff	4.045549
Security staff	3.188518
Cooking staff	2.820741
Cleaning staff	2.207141
Private service staff	1.257329
Low-skill Laborers	0.992855
Waiters/barmen staff	0.638469
Secretaries	0.618382
Realty agents	0.354864
HR staff	0.266866
IT staff	0.244388

Defaulter Percentage within OCCUPATION_TYPE

	TARGET
OCCUPATION_TYPE	
Low-skill Laborers	17.292871
Waiters/barmen staff	11.385768
Drivers	11.367577
Security staff	10.754462
Laborers	10.608441
Cooking staff	10.512038
Cleaning staff	9.642470
Sales staff	9.635105
Realty agents	7.951482
Secretaries	7.115236
Medicine staff	6.691098
IT staff	6.653620
Private service staff	6.542412
Core staff	6.331157
Managers	6.251776
High skill tech staff	6.180623
HR staff	6.093190
Accountants	4.866983

More than (40% combined) of the applicants are either Laborers or Sales Staff (as seen in the univariate analysis) as compared to Low Skilled Laborers who are <1%, but applicants who are Low Skilled Laborers, have the highest defaulting percent of 17.2% in it's category

Recommendation : Going by the proportion of defaulters to applicants, it's safer to give a loan to applicants who are Laborers or Sales Staff as they take up more than 40% applications, yet their defaulter percentage is only about 20% as compared to 17.2% amongst Low Skilled Laborers who only take up <1% of applications. Accountants are the least defaulters with about 4.86% default %

BIVARIATE ANALYSIS
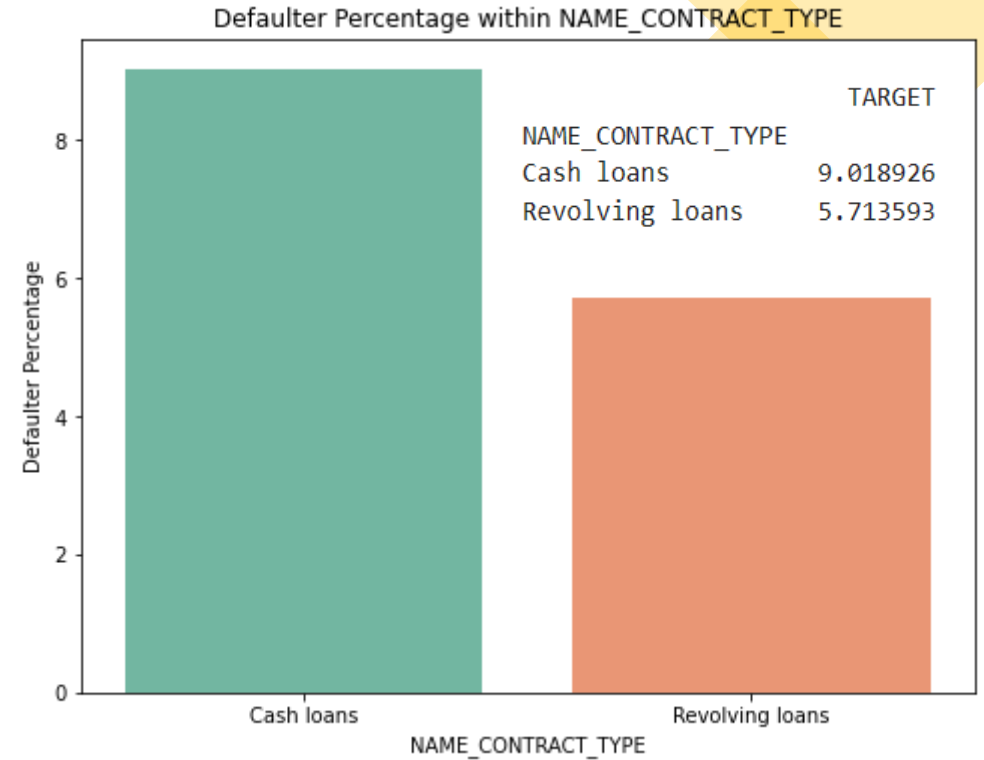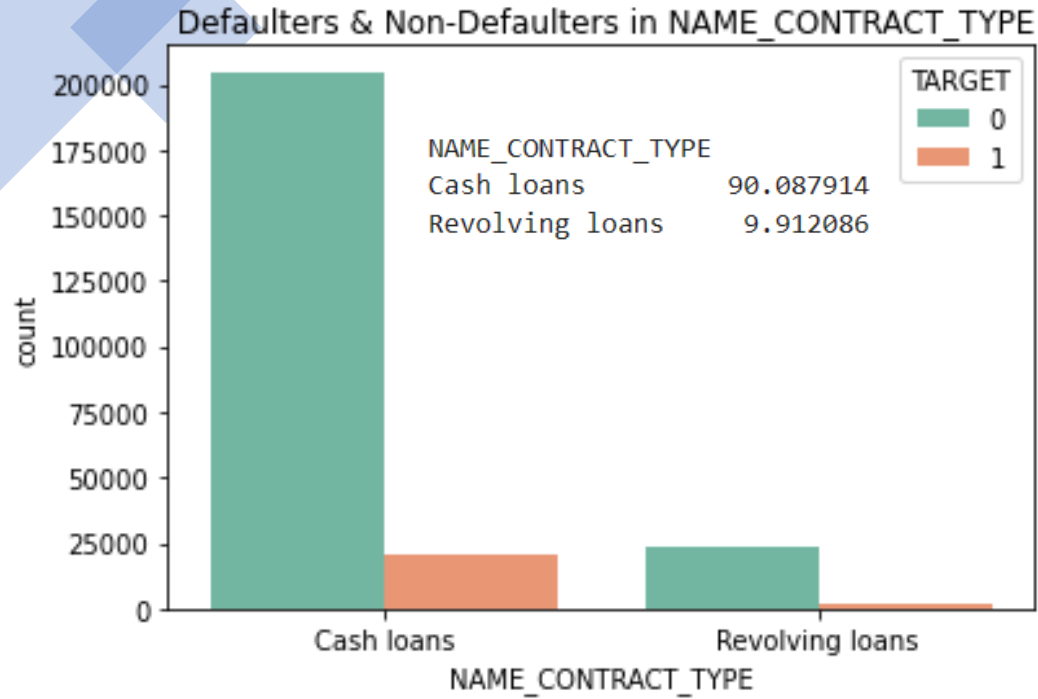
More than 90% of the applications are either Cash Loan (as seen in the univariate analysis) as compared to Revolving Loans who are <10%, but the defaulting rate is 9% for Cash Loans and close to it, 6% for Revolving Loans

Recommendation : Going by the proportion of defaulters to applicants, it's safer to give Cash Loans as the default rate for its proportion to number of applications is low.

BIVARIATE ANALYSIS
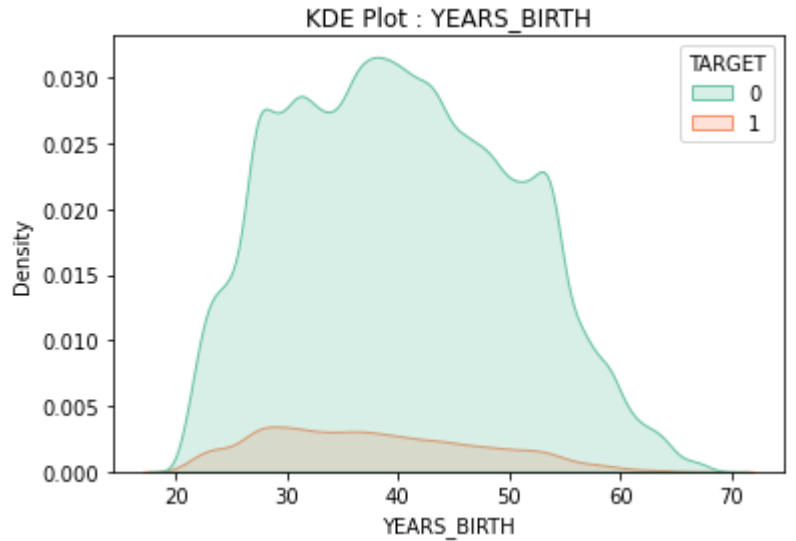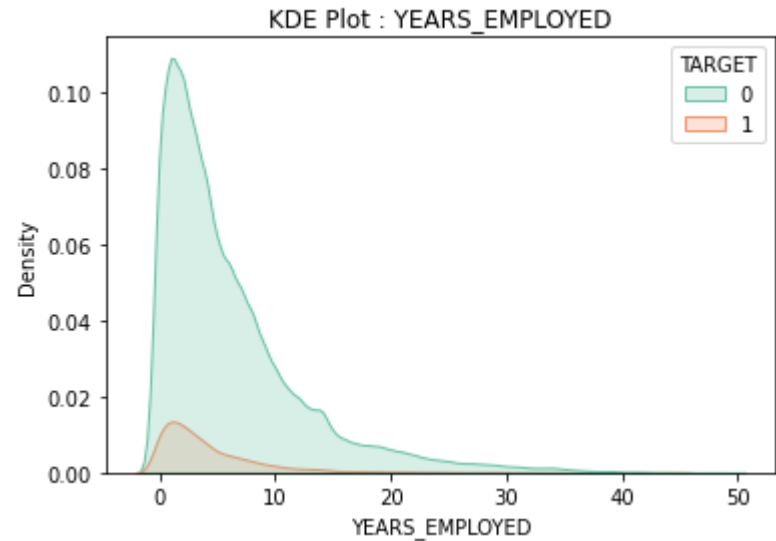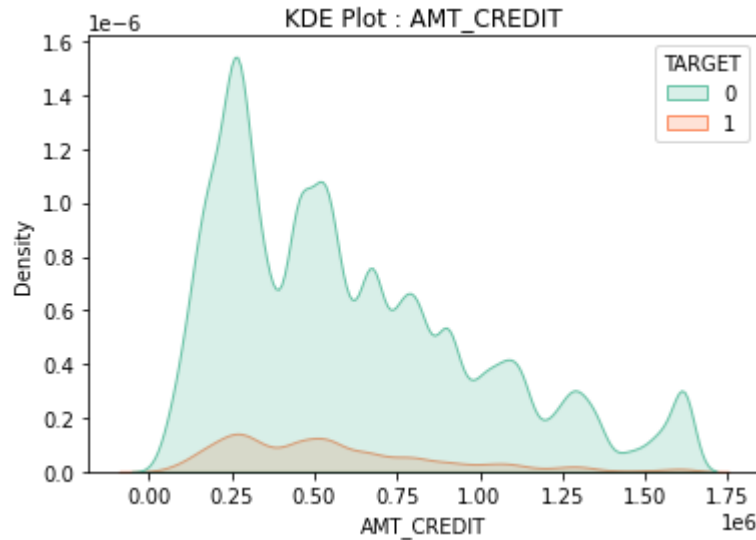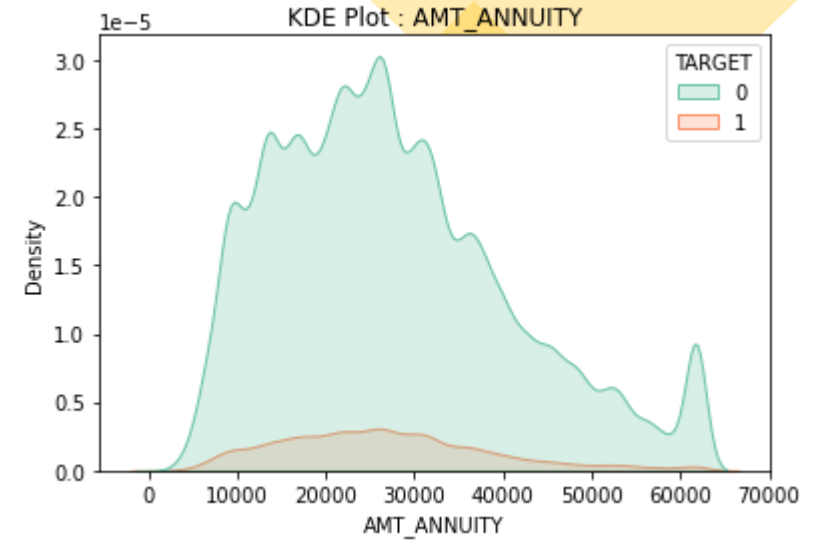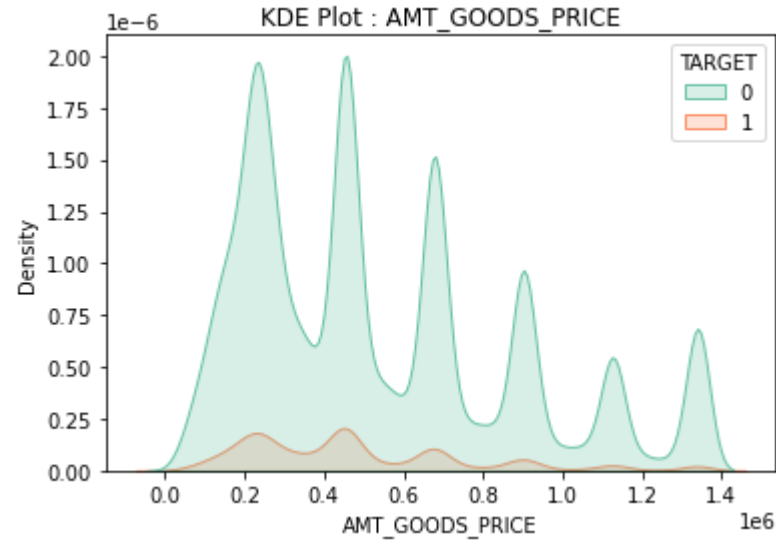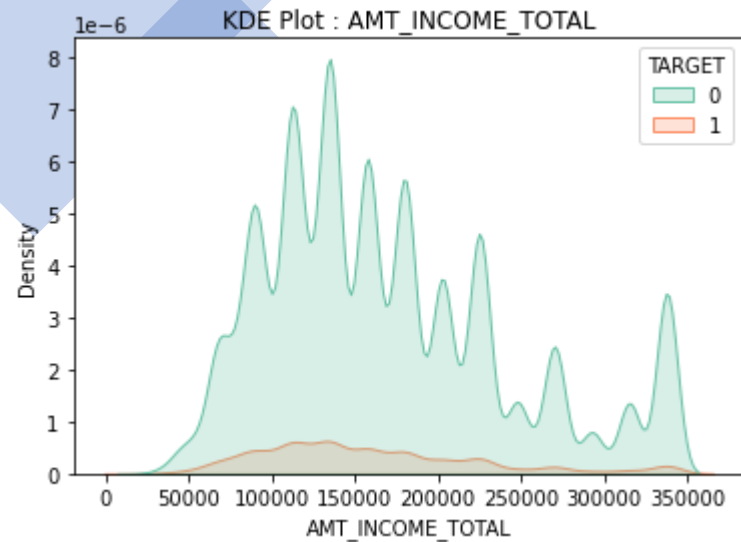
- There is no range of values in any of the plots where only non-defaulters are present
- Income Amount and Goods Price have peaks and valleys at various points : signifying many high values and few low values
- The Years Employed are significantly higher in the range 0-5 years and then reduces significantly.
  - Recommendation : People with years employed between 0-5 have highest non-defaulters, they can be targeted
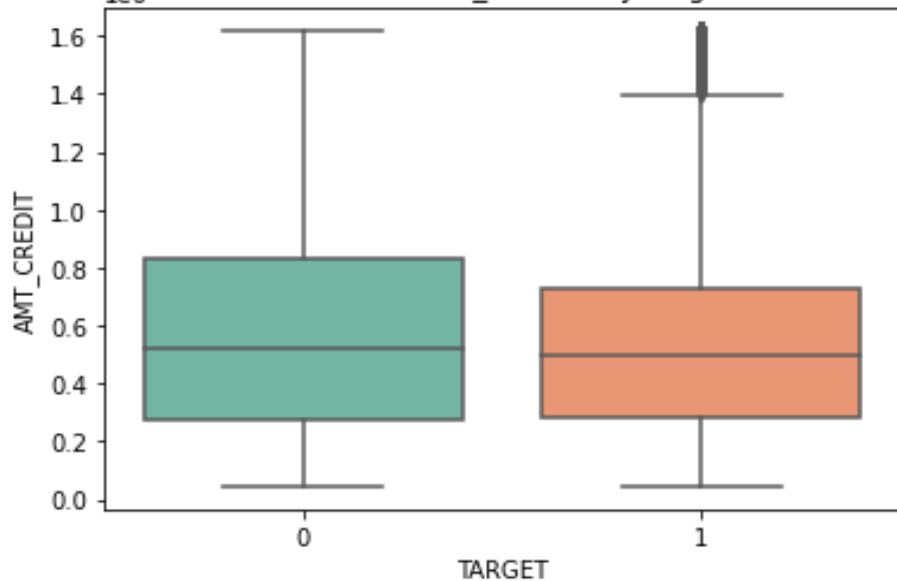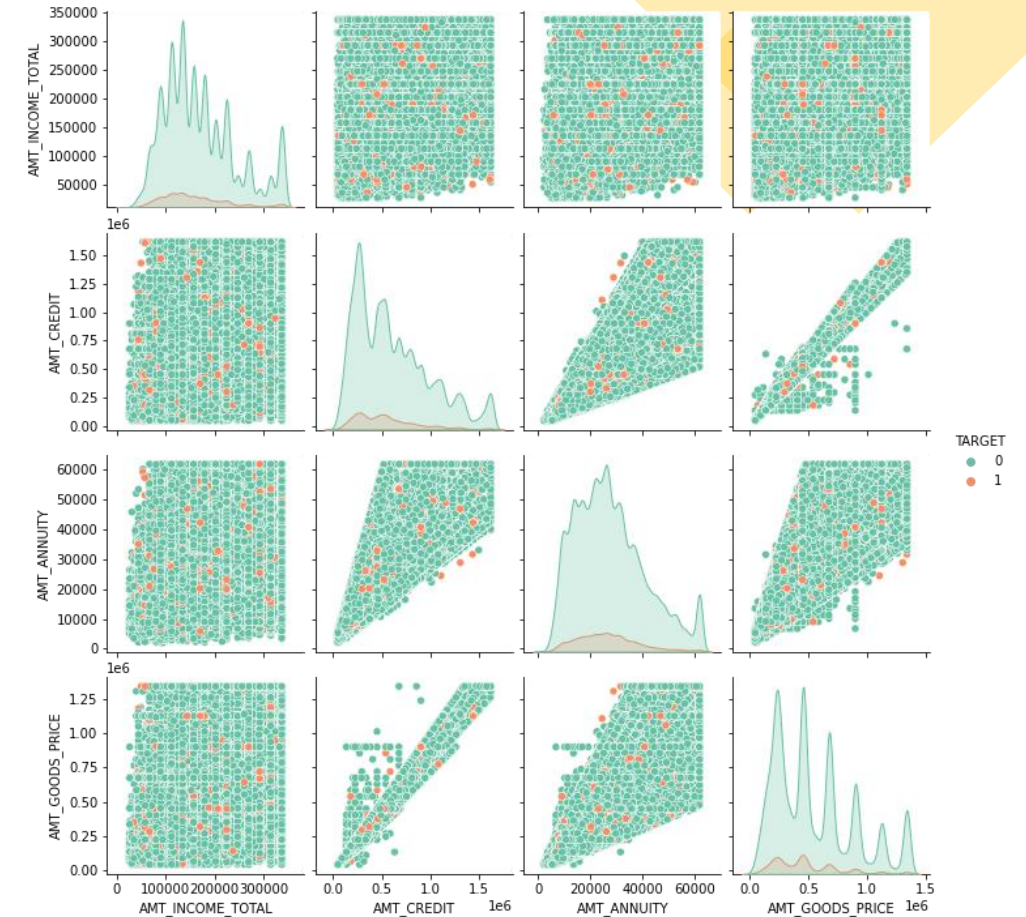
BIVARIATE ANALYSIS
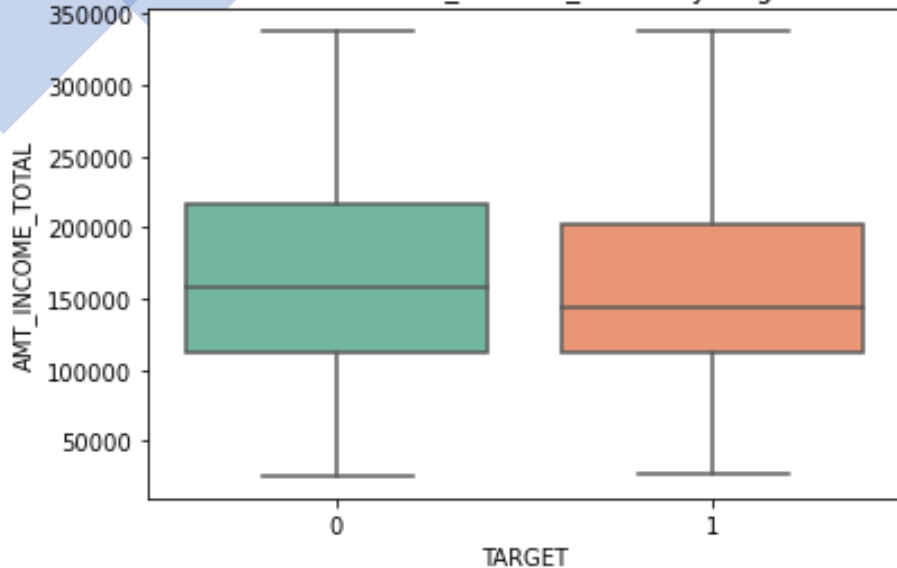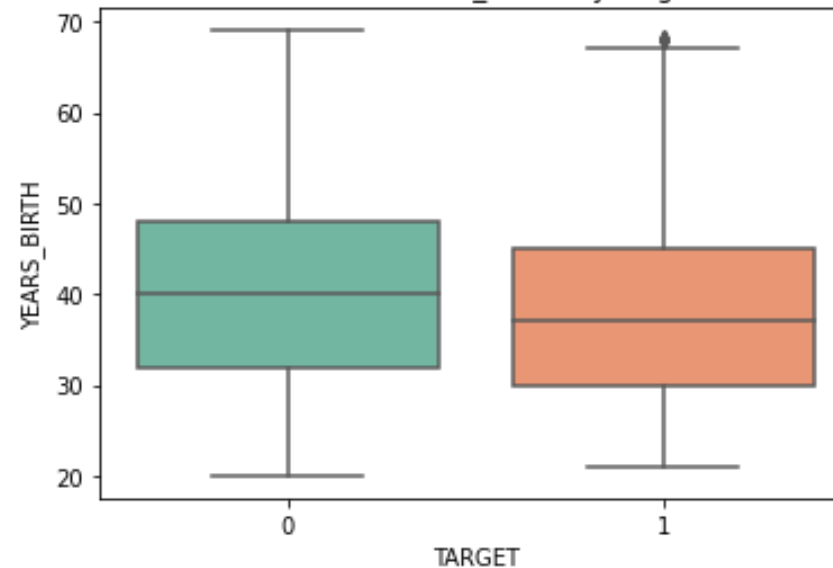
- The median income of defaulters is lower than non-defaulters and is around 1.4lakhs

- The median age of defaulters is lower than non-defaulters and is around 37-38 years of age. This is expected as they have lower incomes since they have been employed for a lesser duration

- The median loan amount appears to be almost the same, although the 75th percentile is around 8 lakhs for non-defaulters and around 7 lakhs for defaulters.

MULTIVARIATE ANALYSIS

DEFAULTERS

	Column 1	Column 2	Correlation
40	AMT_GOODS_PRICE	AMT_CREDIT	0.981745
83	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956380
52	CNT_FAM_MEMBERS	CNT_CHILDREN	0.893874
167	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868951
111	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.846245
139	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.767803
41	AMT_GOODS_PRICE	AMT_ANNUITY	0.756892
27	AMT_ANNUITY	AMT_CREDIT	0.755463

Credit Amount has the highest correlation with Goods Price of   0.98
Credit amount also has a high correlation with Annuity Amount of   0.75
Good Price also has a high correlation with Annuity Amount of   0.75

NON - DEFAULTERS

	Column 1	Column 2	Correlation
40	AMT_GOODS_PRICE	AMT_CREDIT	0.985222
83	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.949221
52	CNT_FAM_MEMBERS	CNT_CHILDREN	0.893379
167	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.861480
111	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.859861
139	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.820925
41	AMT_GOODS_PRICE	AMT_ANNUITY	0.787731
27	AMT_ANNUITY	AMT_CREDIT	0.785454
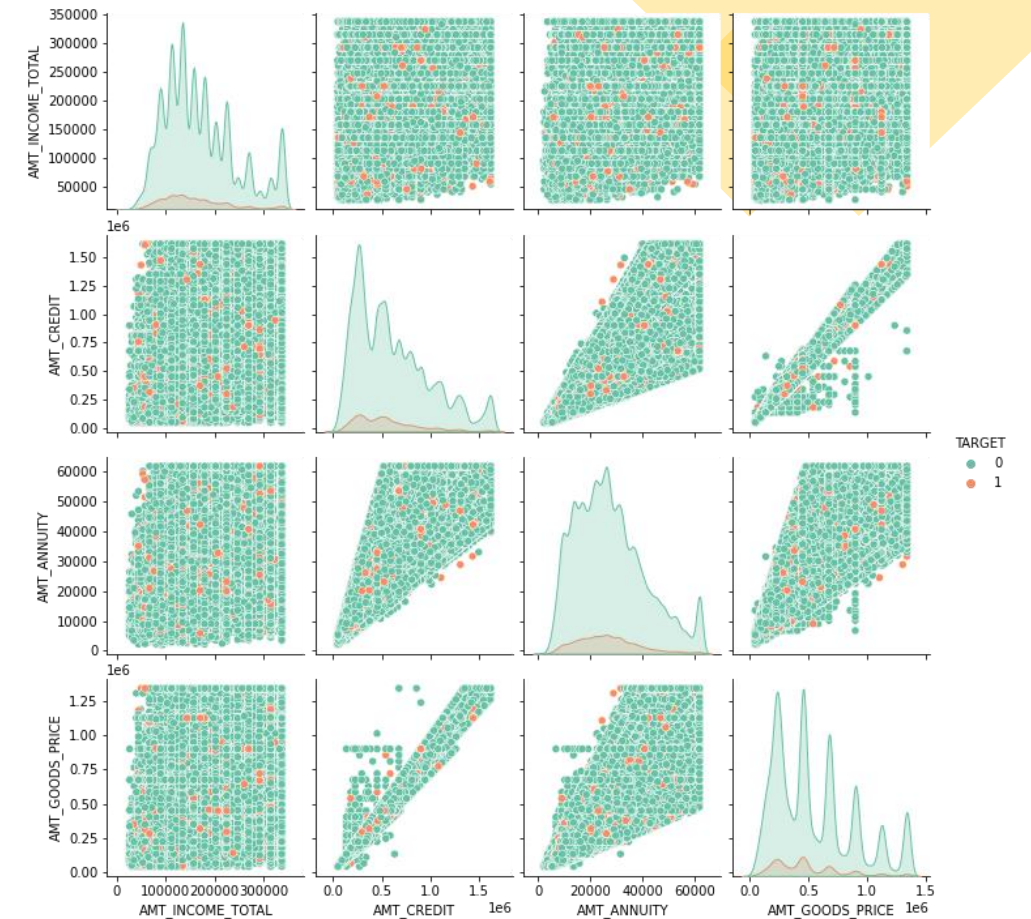
Credit Amount has the highest correlation with Goods Price of   0.98
Credit amount also has a high correlation with Annuity Amount of   0.78
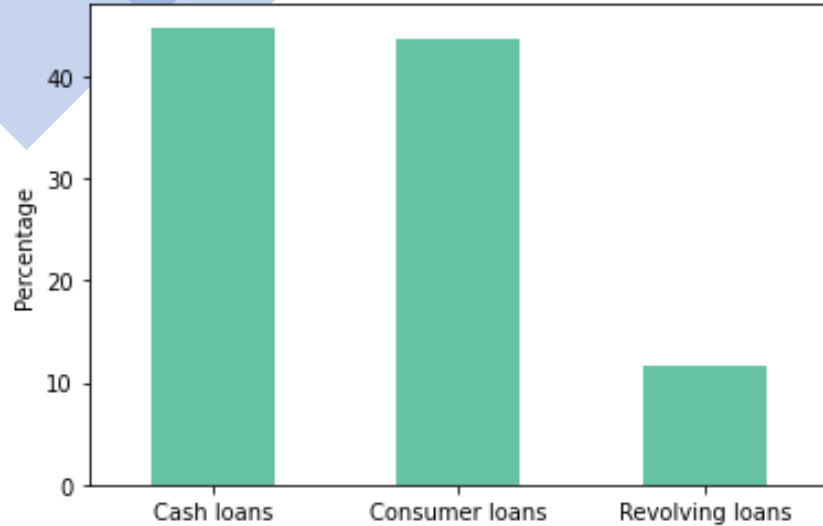Good Price also has a high correlation with Annuity Amount of   0.78
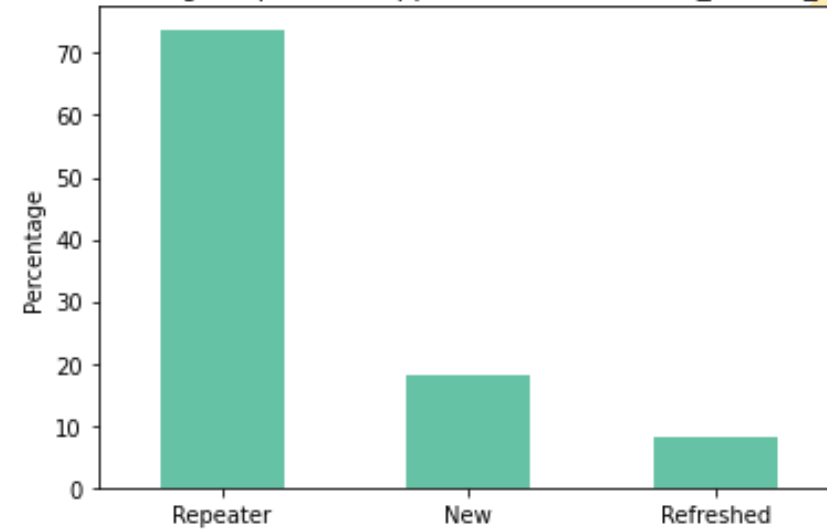
PREVIOUS APPLICATION DATASET

UNIVARIATE ANALYSIS

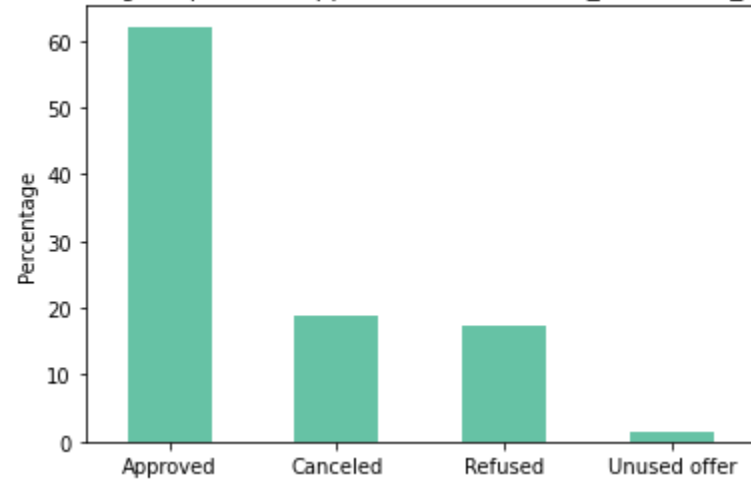Percentage of previous applicants across NAME_CONTRACT_TYPE

Cash Loans and Consumer Loans both constitute almost 45% loan types

Percentage of previous applicants across NAME_CLIENT_TYPE
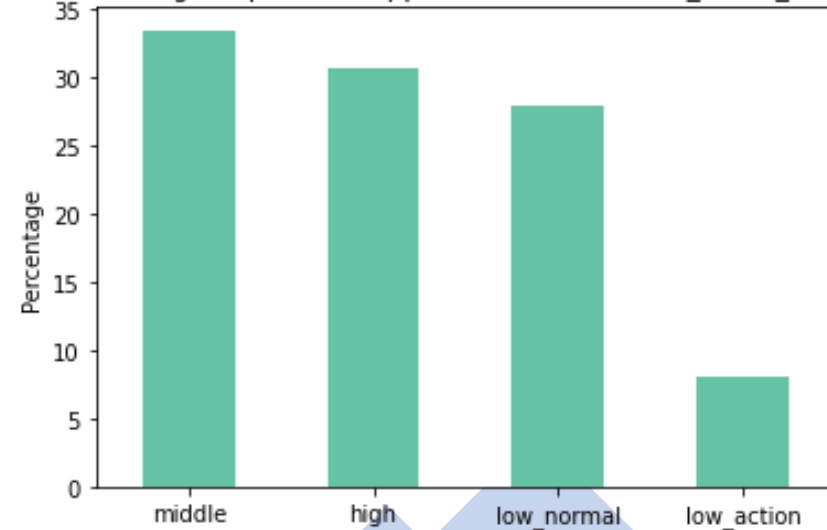
More than 70% of previous applications were Repeaters. New applicants were around 20%

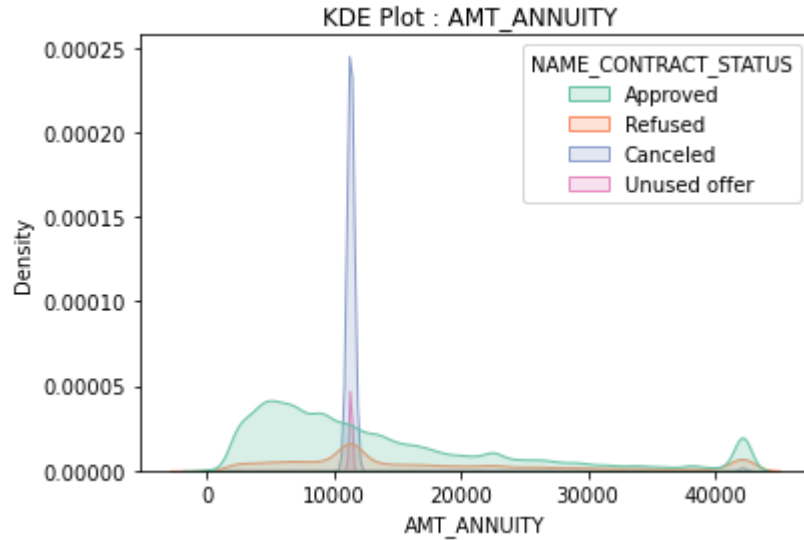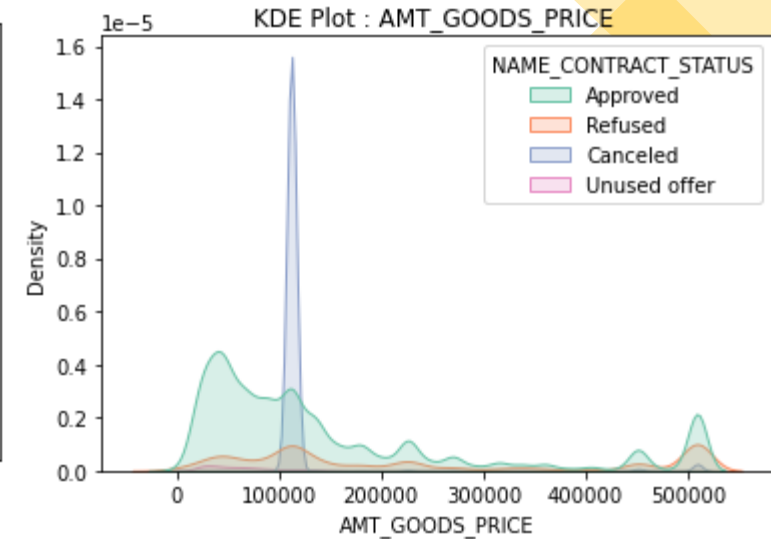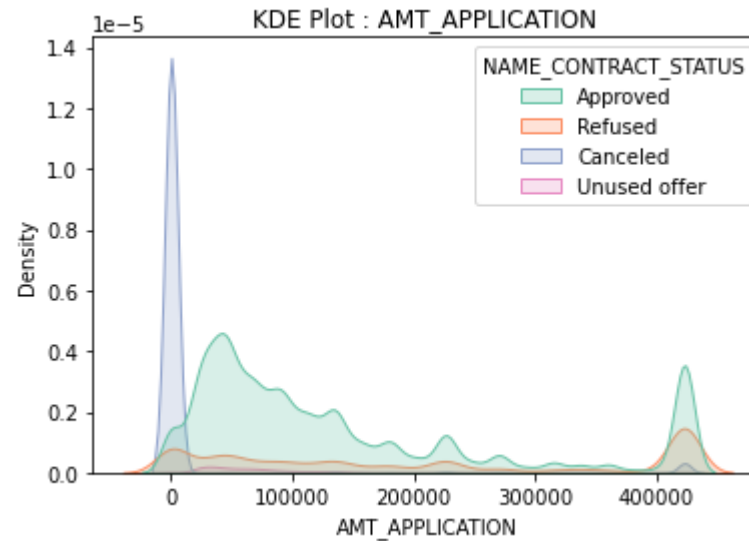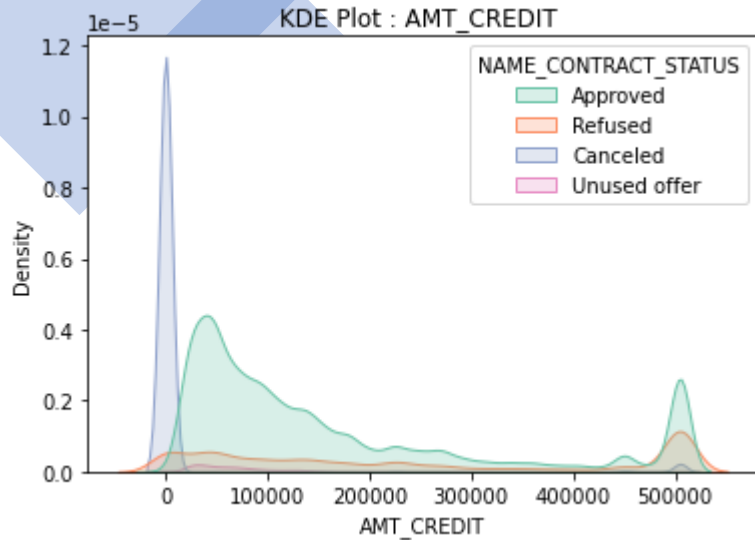Percentage of previous applicants across NAME_CONTRACT_STATUS

More than 60% of previous applications were Approved

Percentage of previous applicants across NAME_YIELD_GROUP

The interest rates of middle, high and low normal categories are almost equal (28-32%) of total previous applications

BIVARIATE ANALYSIS

- The previous applications that were Cancelled appear in only a small range of amounts across all categories, as seen in the graphs
- The Approved applications are distributed fairly across all amounts across all categories, although the approvals become less towards the 4 Lakh range and then increase again for 5 Lakh
- For Amount credited and Amount requested (AMT_CREDIT, AMT_APPLICATION), Refusals appear more at the 5 Lakh and 4.5 Lakh peaks respectively
- For goods price, the Refusals are more at two peaks, 1.2 Lakhs and just over 5 Lakhs
- For annuity amount, Refusals are more that two peaks, 1.2 Lakhs and just over 4 Lakhs
- Unused offers appear only for a small range in Annuity amount, mainly at around 1.2 Lakhs

MULTIVARIATE ANALYSIS

Column 1	Column 2	Correlation
AMT_CREDIT	AMT_APPLICATION	0.941141
AMT_GOODS_PRICE	AMT_APPLICATION	0.941061
AMT_GOODS_PRICE	AMT_CREDIT	0.923592
AMT_ANNUITY	AMT_GOODS_PRICE	0.857584
AMT_ANNUITY	AMT_CREDIT	0.823068
AMT_APPLICATION	AMT_ANNUITY	0.798669
CNT_PAYMENT	AMT_GOODS_PRICE	0.691801
CNT_PAYMENT	AMT_APPLICATION	0.669099
CNT_PAYMENT	AMT_CREDIT	0.637245
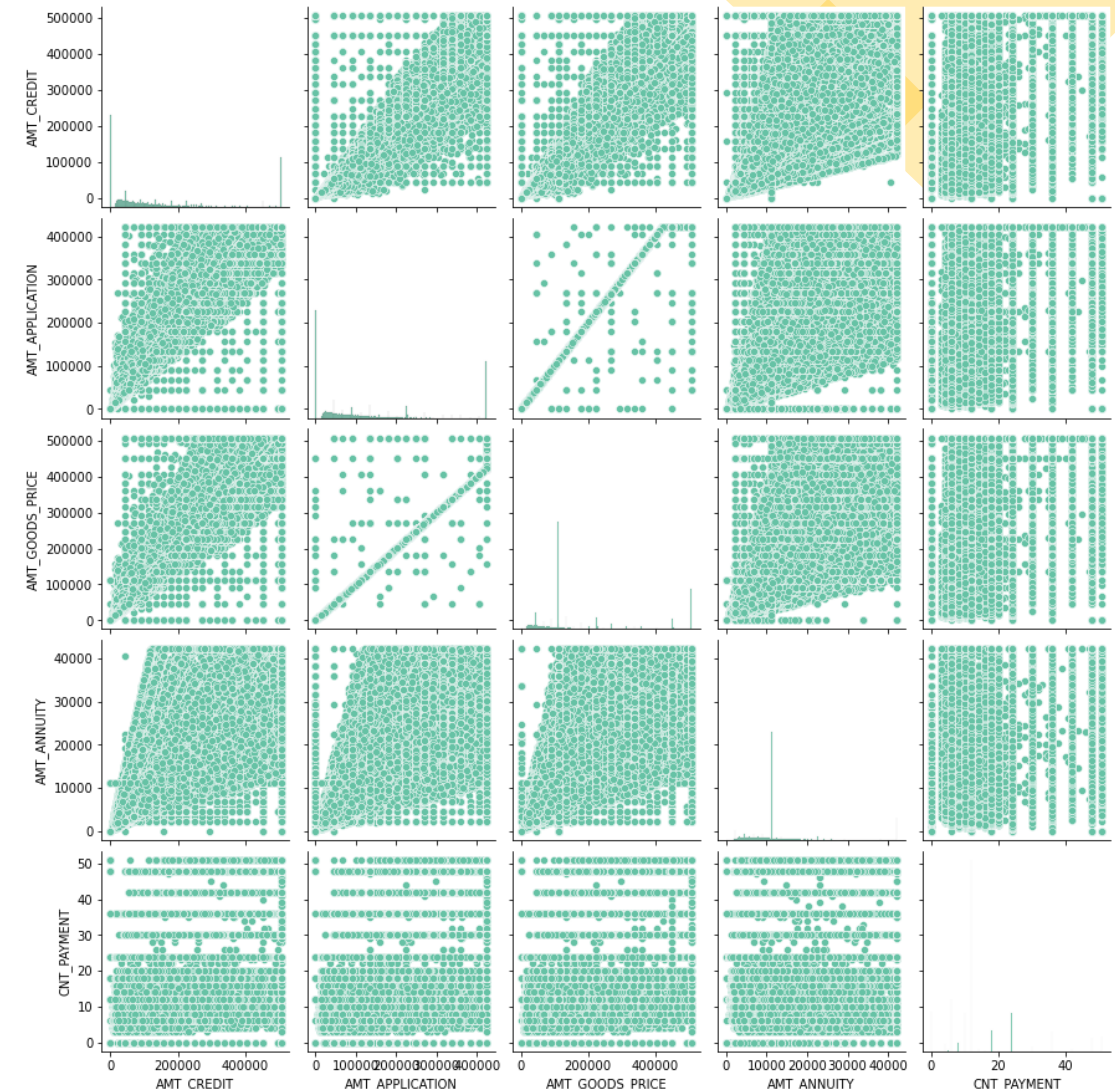
Credit Amount has the highest correlation with Application Amount of 0.94

Goods Price also has a highest correlation with Application Amount of 0.94

MERGED DATASET
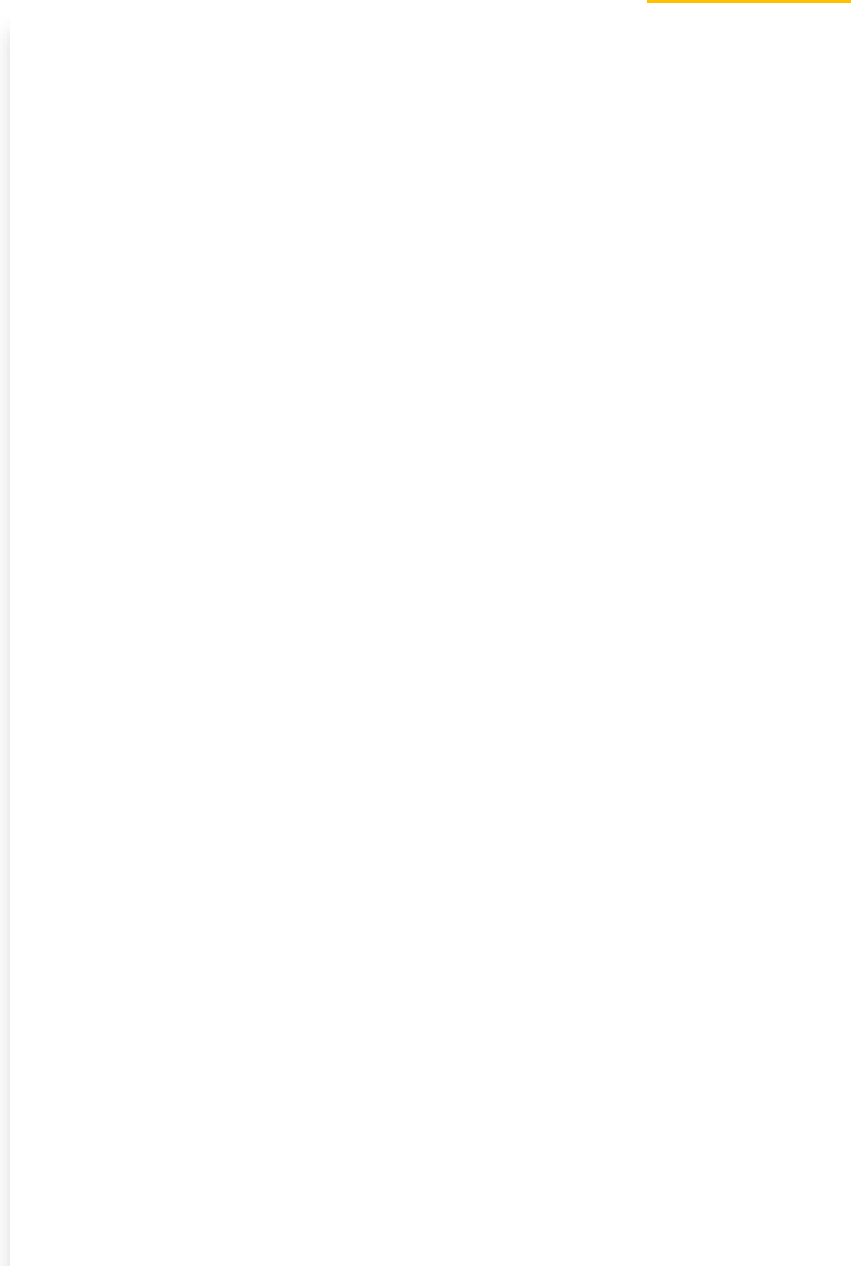
BIVARIATE ANALYSIS

Distribution of CODE_GENDER by CONTRACT STATUS
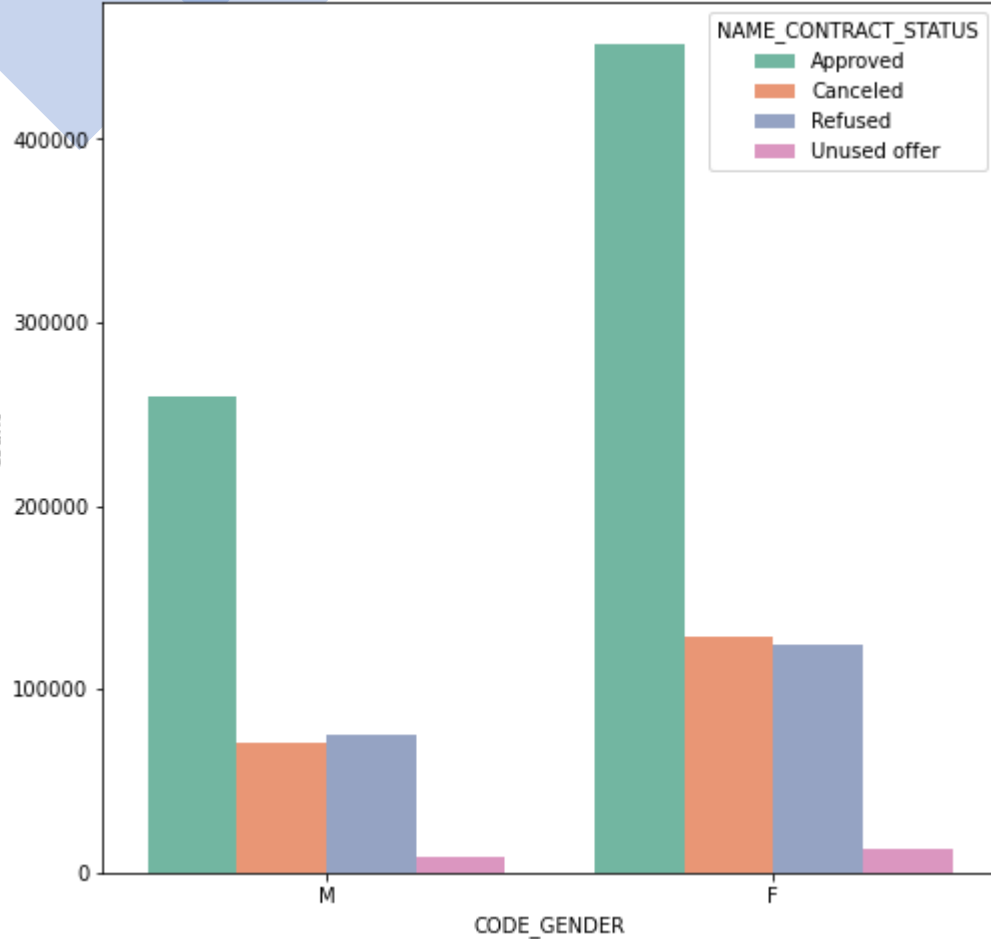
	CODE_GENDER	NAME_CONTRACT_STATUS	Percent
0	F	Approved	62.974573
1	F	Canceled	17.934122
2	F	Refused	17.352883
3	F	Unused offer	1.738423
4	M	Approved	62.617554
5	M	Canceled	17.172322
6	M	Refused	18.225356
7	M	Unused offer	1.984768

The contract statuses of previous applications is fairly equal across both Genders

Distribution of NAME_INCOME_TYPE by CONTRACT STATUS
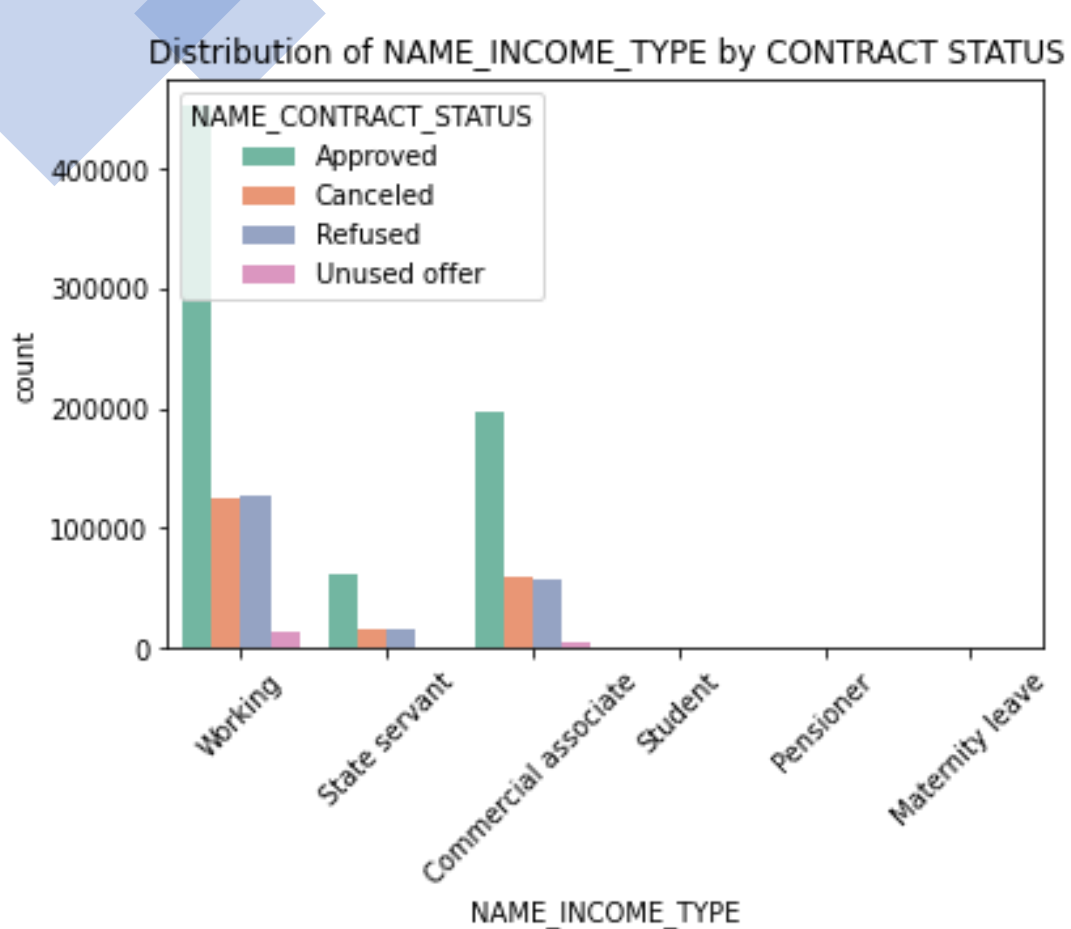
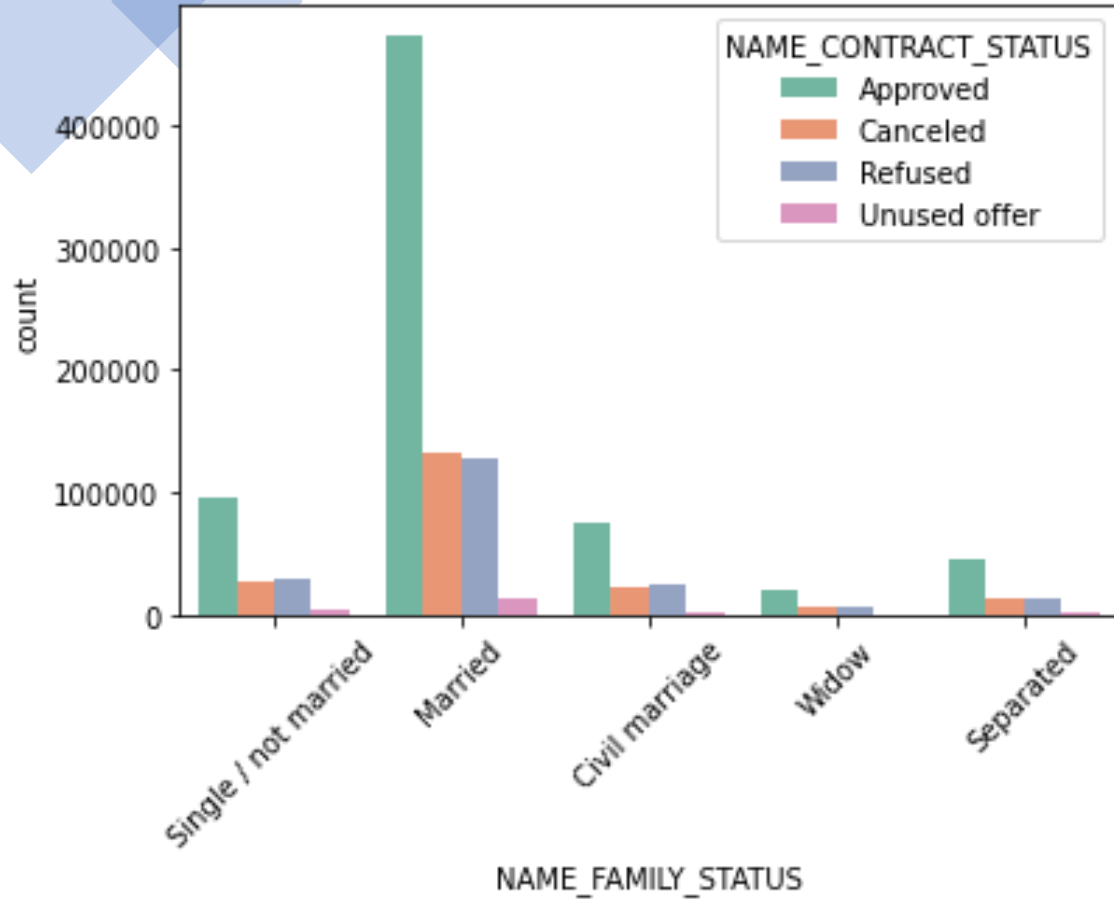|	NAME_INCOME_TYPE	NAME_CONTRACT_STATUS	Percent
0	Commercial associate	Approved	61.801785
1	Commercial associate	Canceled	18.555111
2	Commercial associate	Refused	18.063305
3	Commercial associate	Unused offer	1.579799
4	Maternity leave	Approved	62.500000
5	Maternity leave	Canceled	12.500000
6	Maternity leave	Refused	18.750000
7	Maternity leave	Unused offer	6.250000
8	Pensioner	Approved	52.054795
9	Pensioner	Canceled	19.178082
10	Pensioner	Refused	28.767123
11	State servant	Approved	65.302901
12	State servant	Canceled	16.580631
13	State servant	Refused	16.507039
14	State servant	Unused offer	1.609428
15	Student	Approved	83.333333
16	Student	Canceled	12.500000
17	Student	Refused	4.166667
18	Working	Approved	62.986276
19	Working	Canceled	17.396567
20	Working	Refused	17.649514
21	Working	Unused offer	1.967643

The Approved rate is the highest for Student at 83%, Refused rate is the lowest for student at 4% and Cancellation rate is lowest at 12.5% in previous applications

Recommendation : Based on previous applications, since Students have the highest approved rate, lowest refused rate and lowest cancellation rate, they are a safe Income Type to target for giving loans

Distribution of NAME_FAMILY_STATUS by CONTRACT STATUS

	NAME_FAMILY_STATUS	NAME_CONTRACT_STATUS	Percent
0	Civil marriage	Approved	60.951179
1	Civil marriage	Canceled	17.770268
2	Civil marriage	Refused	19.773606
3	Civil marriage	Unused offer	1.504947
4	Married	Approved	63.495145
5	Married	Canceled	17.724049
6	Married	Refused	17.008790
7	Married	Unused offer	1.772017
8	Separated	Approved	62.435276
9	Separated	Canceled	17.358443
10	Separated	Refused	18.214465
11	Separated	Unused offer	1.991816
12	Single / not married	Approved	61.672920
13	Single / not married	Canceled	17.234380
14	Single / not married	Refused	18.677568
15	Single / not married	Unused offer	2.415131
16	Widow	Approved	61.610365
17	Widow	Canceled	18.301485
18	Widow	Refused	18.893096
19	Widow	Unused offer	1.195054

The contract statuses of previous applications is fairly equal across all family statuses

Distribution of NAME_EDUCATION_TYPE by CONTRACT STATUS
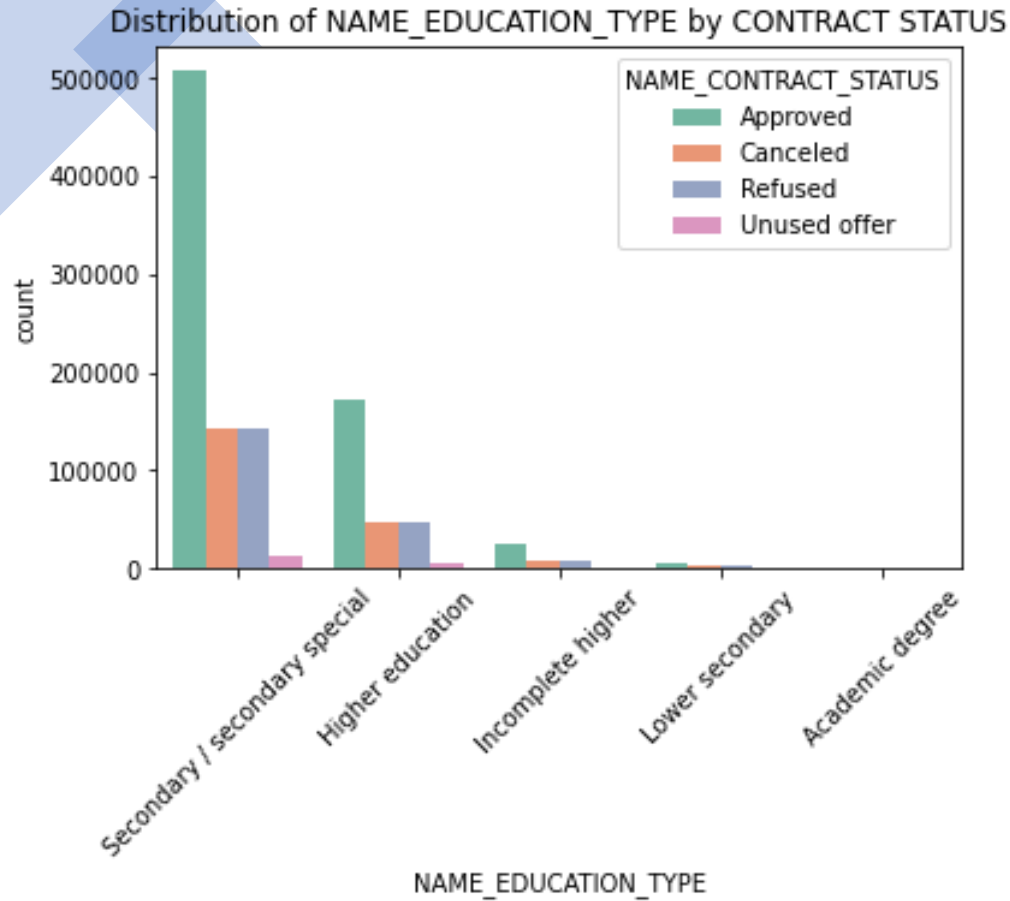
	NAME_EDUCATION_TYPE	NAME_CONTRACT_STATUS	Percent
0	Academic degree	Approved	71.602434
1	Academic degree	Canceled	9.330629
2	Academic degree	Refused	17.849899
3	Academic degree	Unused offer	1.217039
4	Higher education	Approved	62.668588
5	Higher education	Canceled	17.612124
6	Higher education	Refused	17.413183
7	Higher education	Unused offer	2.306106
8	Incomplete higher	Approved	61.245617
9	Incomplete higher	Canceled	17.269971
10	Incomplete higher	Refused	19.200701
11	Incomplete higher	Unused offer	2.283711
12	Lower secondary	Approved	62.417431
13	Lower secondary	Canceled	17.923691
14	Lower secondary	Refused	18.406783
15	Lower secondary	Unused offer	1.252095
16	Secondary / secondary special	Approved	62.987361
17	Secondary / secondary special	Canceled	17.692184
18	Secondary / secondary special	Refused	17.670330
19	Secondary / secondary special	Unused offer	1.650125

Applicants with Academic degrees have the highest Approval rate of 71% and the lowest cancelled rate of 9% amongst previous applications
Recommendation : Applicants with Academic degrees cancel the least and due to high Approved rate in previous applications, they can be targeted

Distribution of TARGET by CONTRACT STATUS
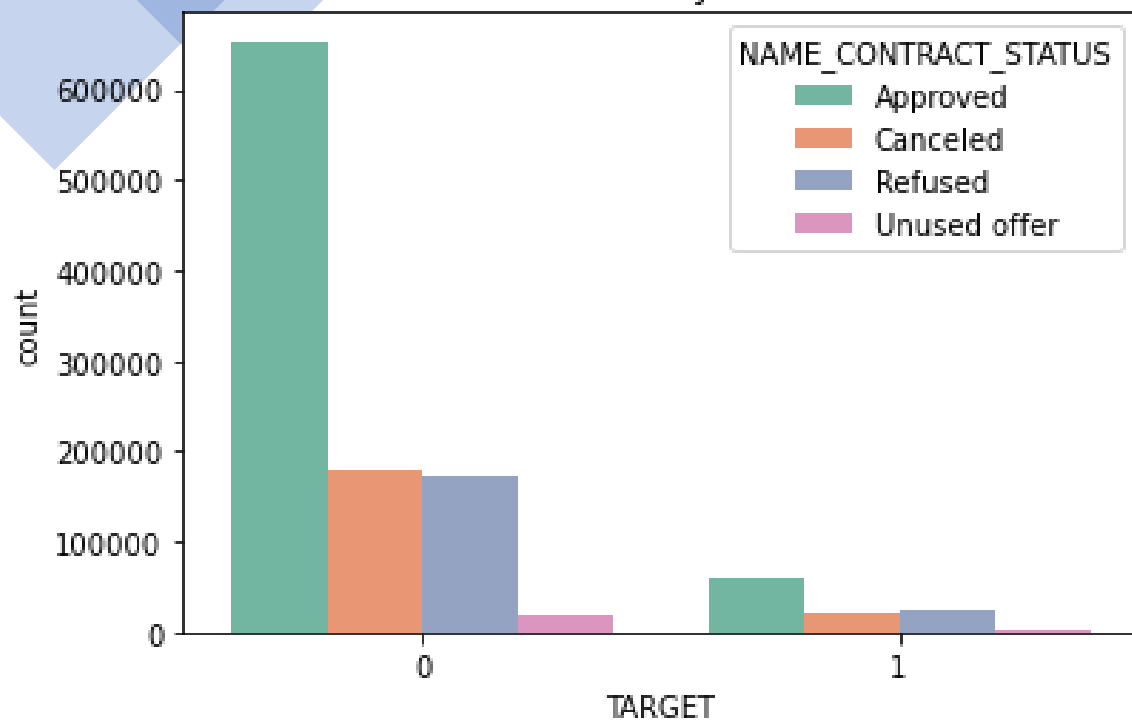
	TARGET	NAME_CONTRACT_STATUS	Percent
0	0	Approved	63.660722
1	0	Canceled	17.515003
2	0	Refused	16.979422
3	0	Unused offer	1.844854
4	1	Approved	54.911965
5	1	Canceled	19.020257
6	1	Refused	24.397955
7	1	Unused offer	1.669822

As expected, the Approval rate for defaulters is lesser at 55% and the Refusal rate is higher at 24%