EDA ASSIGNMENT

- DEVANSHI GUPTA

INDEX

FLOWCHART OF ANALYSIS

IMPORTING LIBRARIES

NumPy, pandas, Matplotlib and seaborn

DATA UNDERSTANDING

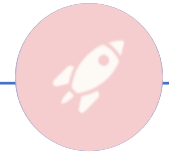
Printing head ,tail, shape, describe and info

PREDICTIVE MODELLING

Univariate categorical, numerical and Bivariate analysis using bar,countplot,corr matrix ,scatterplot, line chart

DRAWING INSIGHTS

Drawing inferences regarding defaulters and repayers to know whom to give loan in future.



CONCLUSION

Drawing overall conclusion to prevent consumers capable of repaying the loan are not rejected.

PROBLEM STATEMENT

Banks face challenges when deciding to provide loans to individuals who have limited or missing credit history. Some people exploit this situation by intentionally becoming defaulters, which means they don't repay their loans on purpose.

When the company receives a loan application, they have to decide whether to approve the loan based on the applicant's profile. This decision involves two types of risks:

1. If the applicant is likely to repay the loan, not approving the loan could result in a business loss for the company because they miss out on a potential opportunity.

2. Conversely, if the applicant is not likely to repay the loan, approving it could lead to a financial loss for the company due to potential defaults or fraudulent activities.

OBJECTIVE

By conducting thorough data analysis, the bank aims to make well-informed decisions to reduce the chances of approving loans for individuals who may not repay while ensuring that deserving applicants are not rejected. The goal is to strike a balance between minimizing risks and providing financial assistance to those capable of repayment.

- ❑ Assumptions: Considering XAP and XAN as missing values

APPROACH AND METHODOLOGY

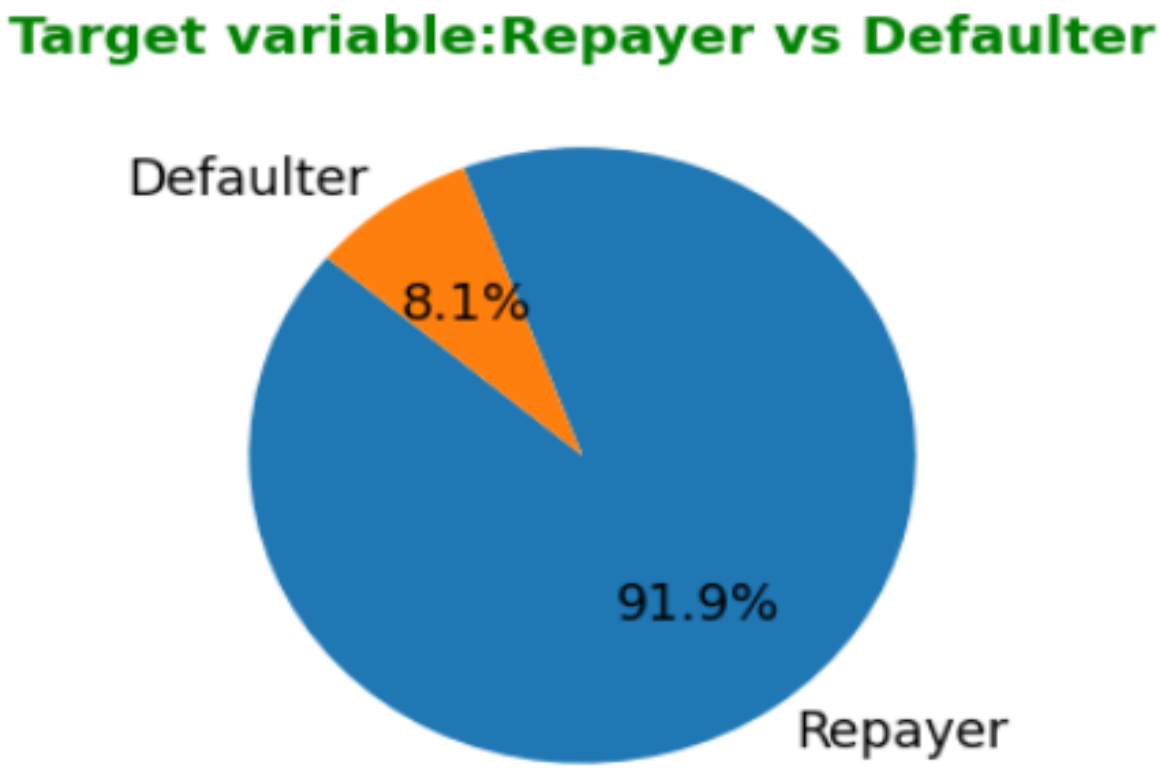
- Loading Application data and Previous application CSV file.
- Look for insights (Describe, Info, Shape)
- Calculating null values as they affect the analysis if present in high percentage.
- Dropping the columns in both dataset with null values more than 40% and carefully dropped as they were of no use for analysis.

1.In Application data, 49 columns are dropped and in Previous application, 11 columns are dropped.

- For categorical columns, null values are replaced by mode.
- For numeric columns, first plotted box plot to check for outliers. If there were outliers which could affect the analysis then replaced null values with median else with mean or either dropped them.

2. Dropped unwanted columns
3. Changed all DAYS columns values from negative to positive.
4. Bifurcated the Application dataset into two parts: Repayer_target and Defaulter_target
5. Binned age, children and income columns.
6. Used univariate and bivariate analysis to draw insights.

DATA IMBALANCE



Observation: 91.9% of people pay back the loan on time whereas 8.1% fail to pay on time.

CATEGORICAL UNIVARIATE ANALYSIS

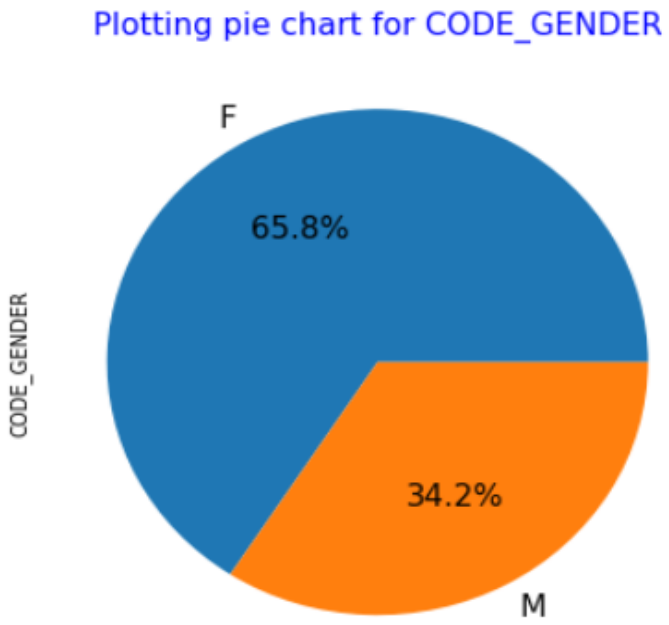
1. Application data

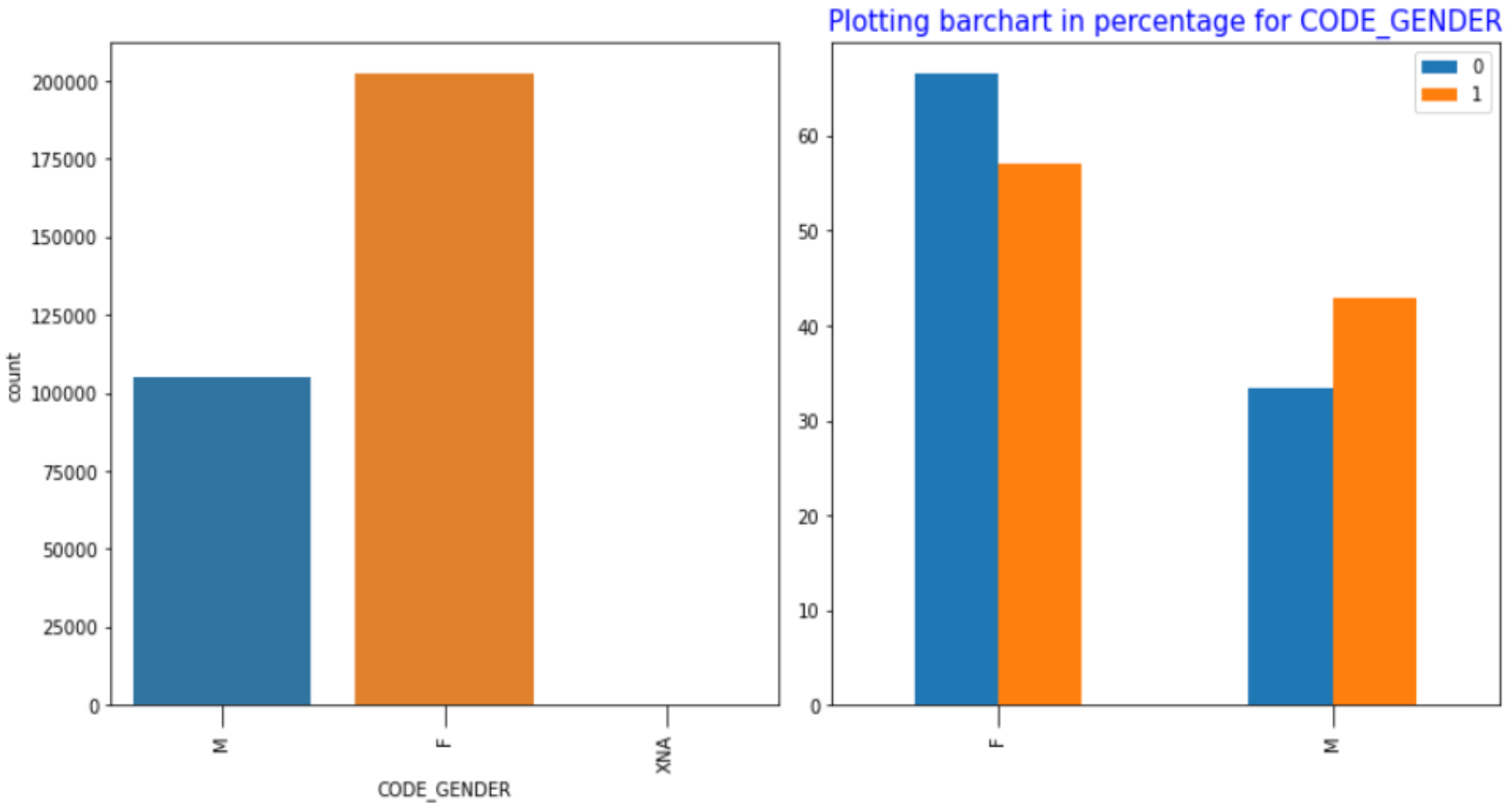
- CODE_GENDER
- NAME_CONTRACT_TYPE
- FLAG_OWN_REALTY
- NAME_INCOME_TYPE
- NAME_EDUCATION_TYPE
- NAME_FAMILY_STATUS
- NAME_HOUSING_TYPE
- OCCUPATION_TYPE

2. Previous Application

- NAME_CONTRACT_STATUS
- NAME_CLIENT_TYPE
- NAME_CASH_LOAN_PURPOSE

CODE_GENDER

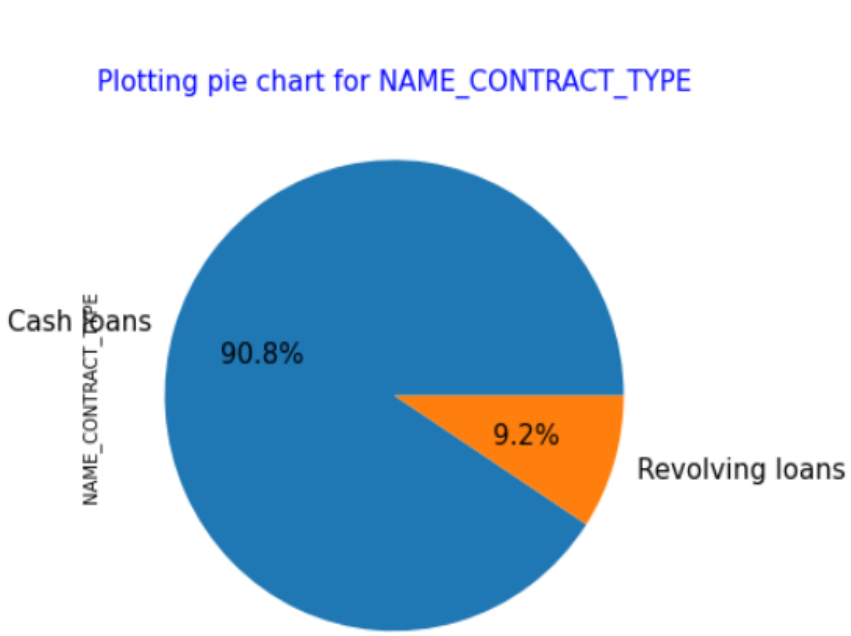


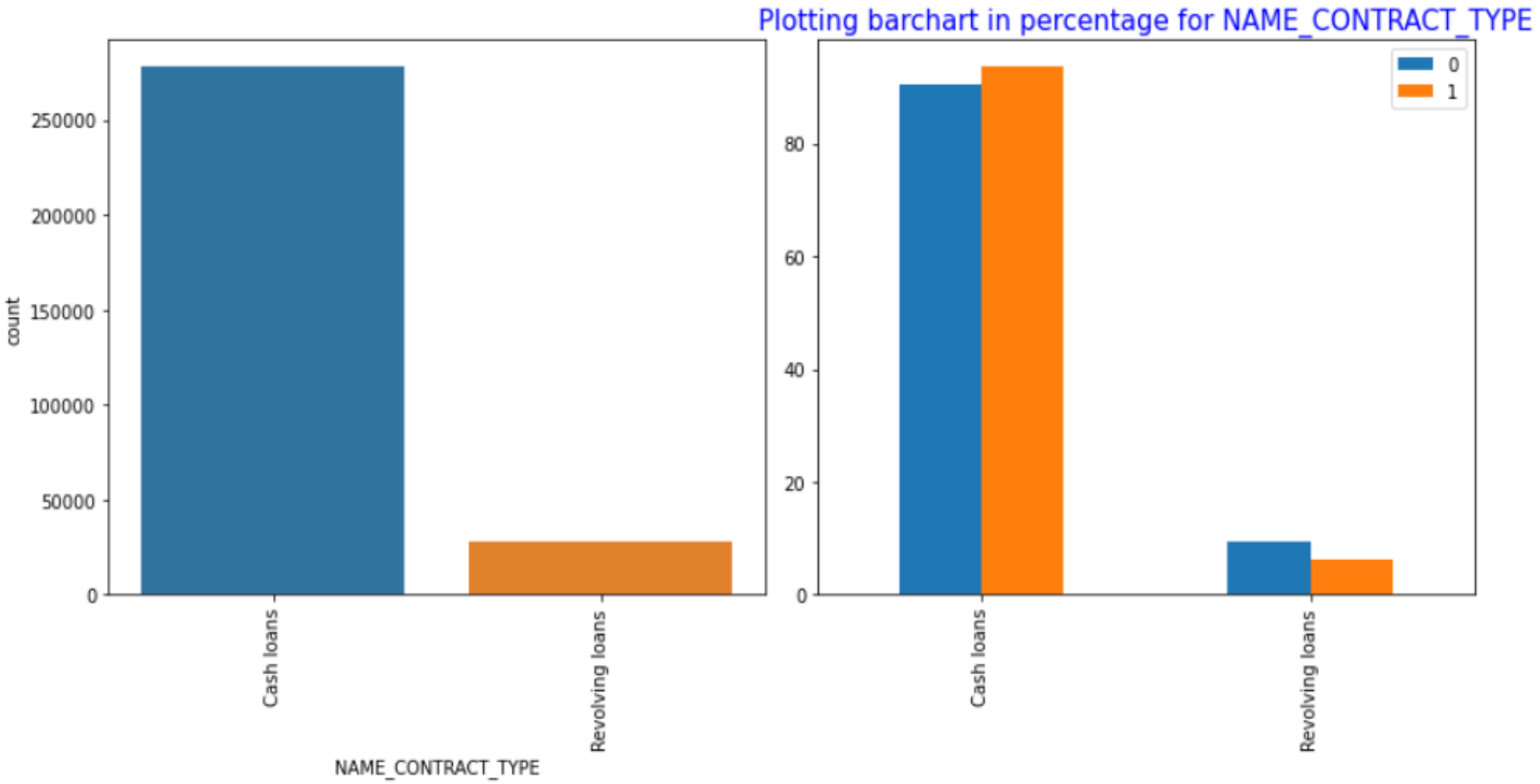


Observation:

- 1.65.8% females and 34.2% take loan.
- 2.The number of females paying on time is almost double of number of men.
- 3.The number of female applying for loan are more than male.
- 4.As men are compared to female , the percentage of defaulter males : 10% are comparatively more to the percentage of males who pay on time
- Hence, accepting loans for females will be less risky than men.

NAME_CONTRACT_TYPE

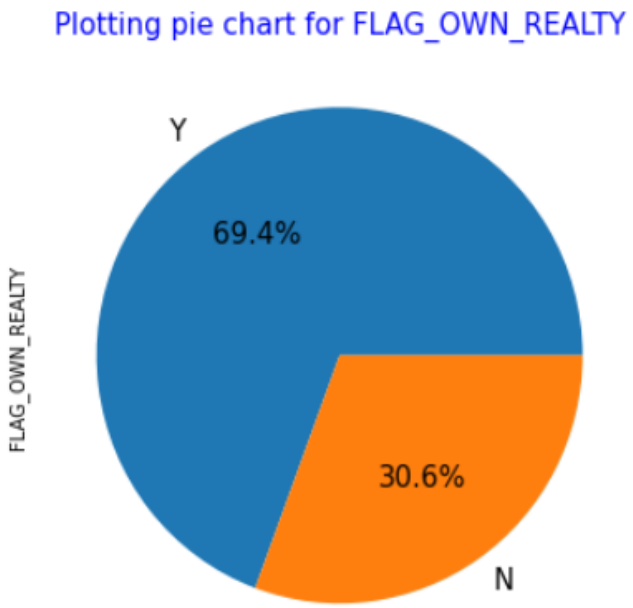


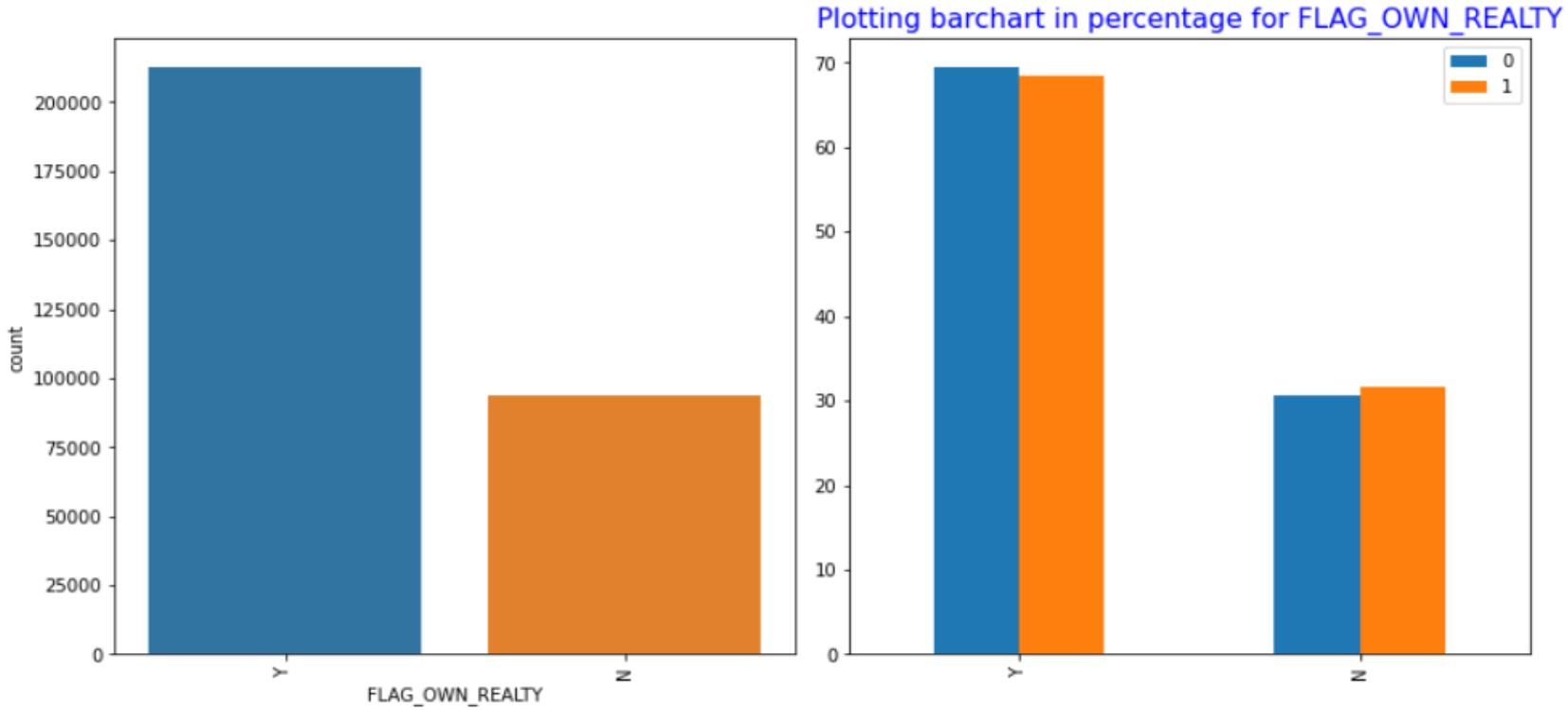


Observation:

- 1.90.8% people take cash loans and 9.2% people take revolving loans.
- 2.More than 80% people taking cash loans fail to pay on time.
- 3.People taking revolving loans are likely to pay on time and the number of defaulter is also less.
- Hence, giving revolving loans will be less risky.

FLAG_OWN_REALTY

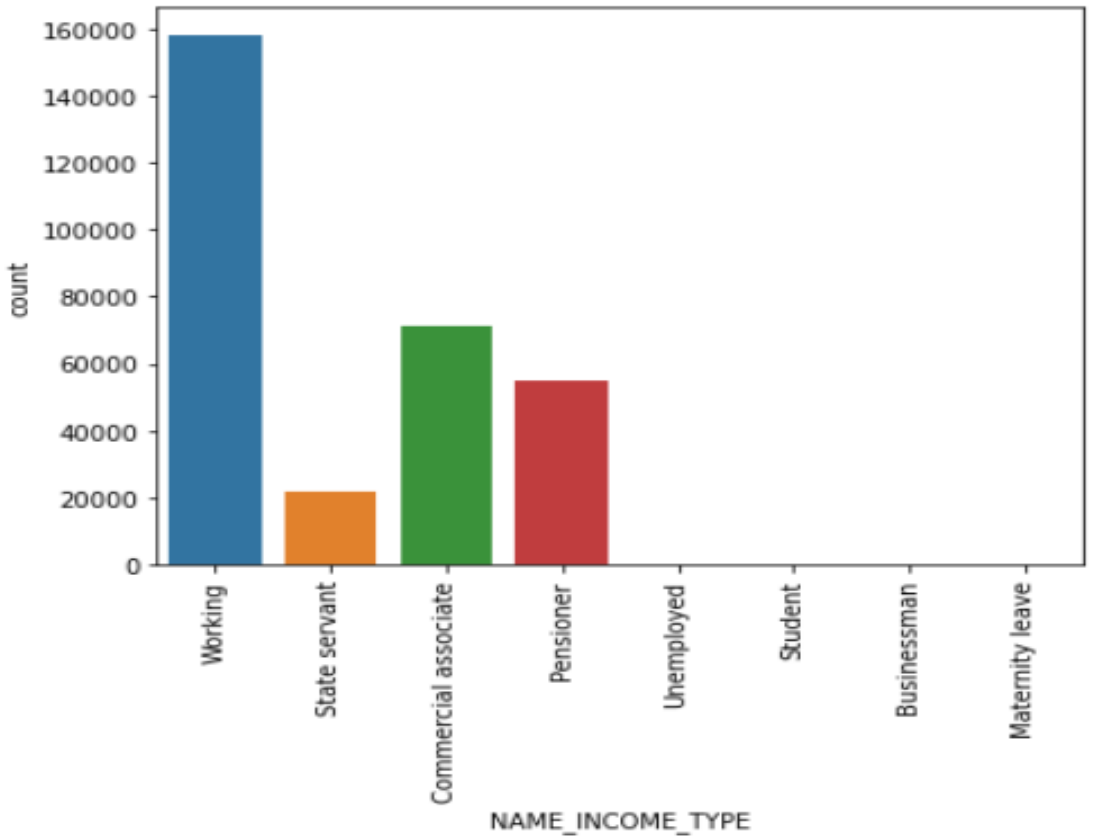


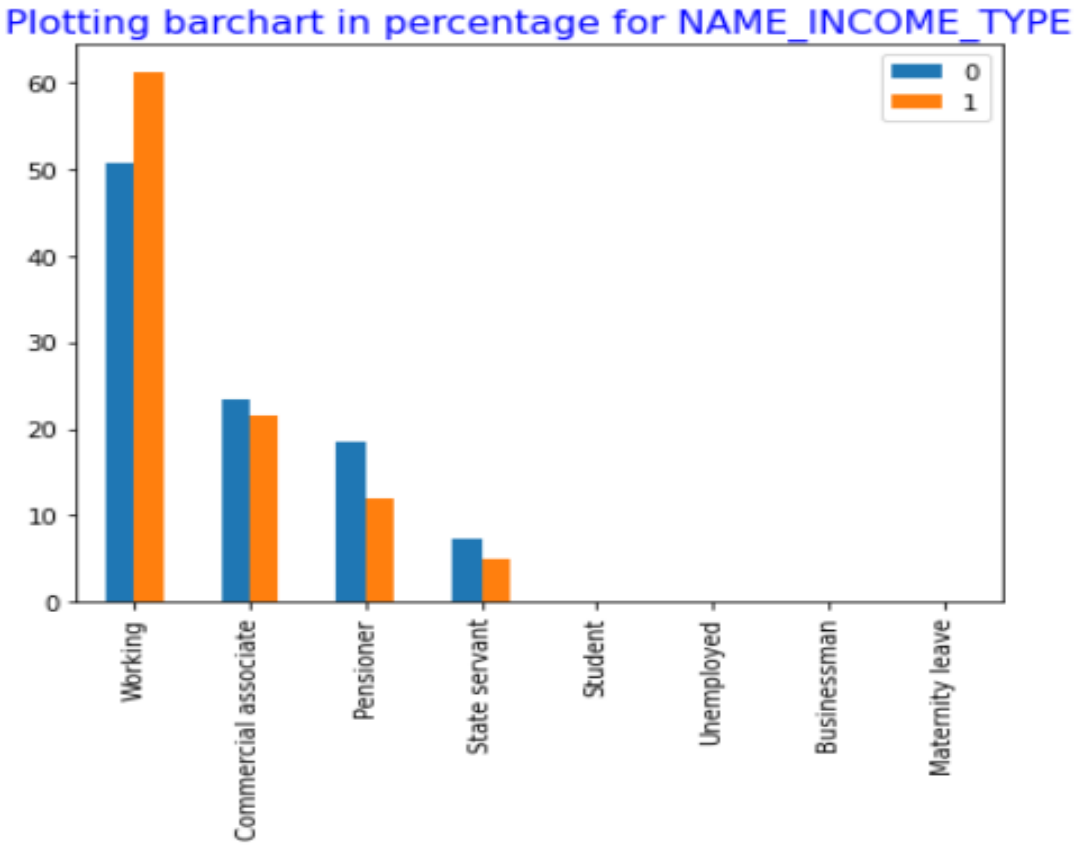


Observation:

- 1.69.4% of people own a house/flat and 30.6% of people don't own a house/flat.
- 2.People having house/flat take more loan.
- 3.People who don't own a house/flat are more likely to face difficulty to pay off loan.

NAME_INCOME_TYPE

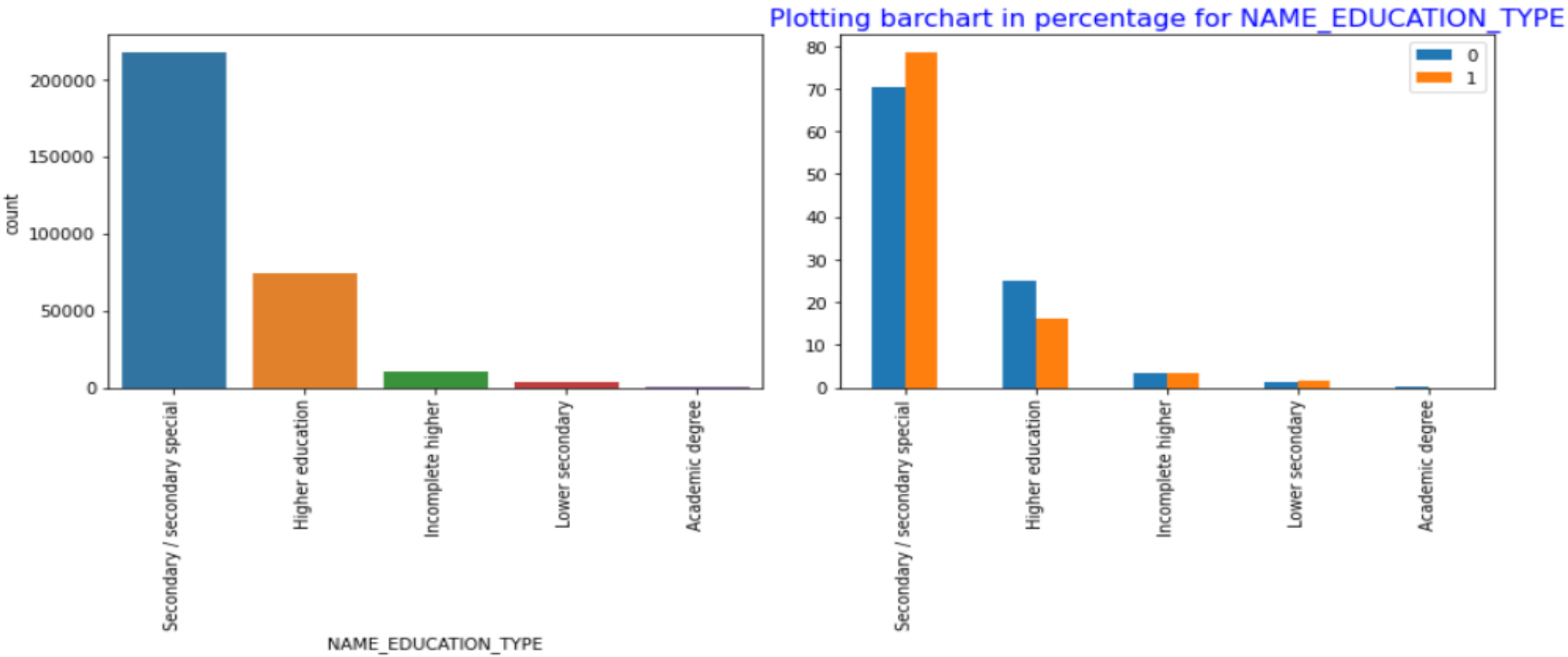




Observation:

- 1.Working,Commercial associate and Pensioner take more loan.
- 2.Commercial associate ,Pensioner and State servant are more likely to pay on time.
- 3.Working class is facing difficulty to pay off loan on time and become defaulter.

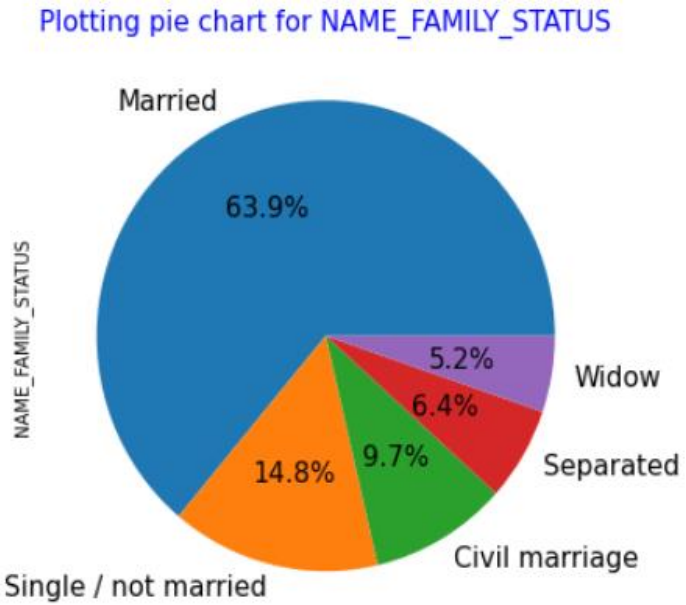
NAME_EDUCATION_TYPE

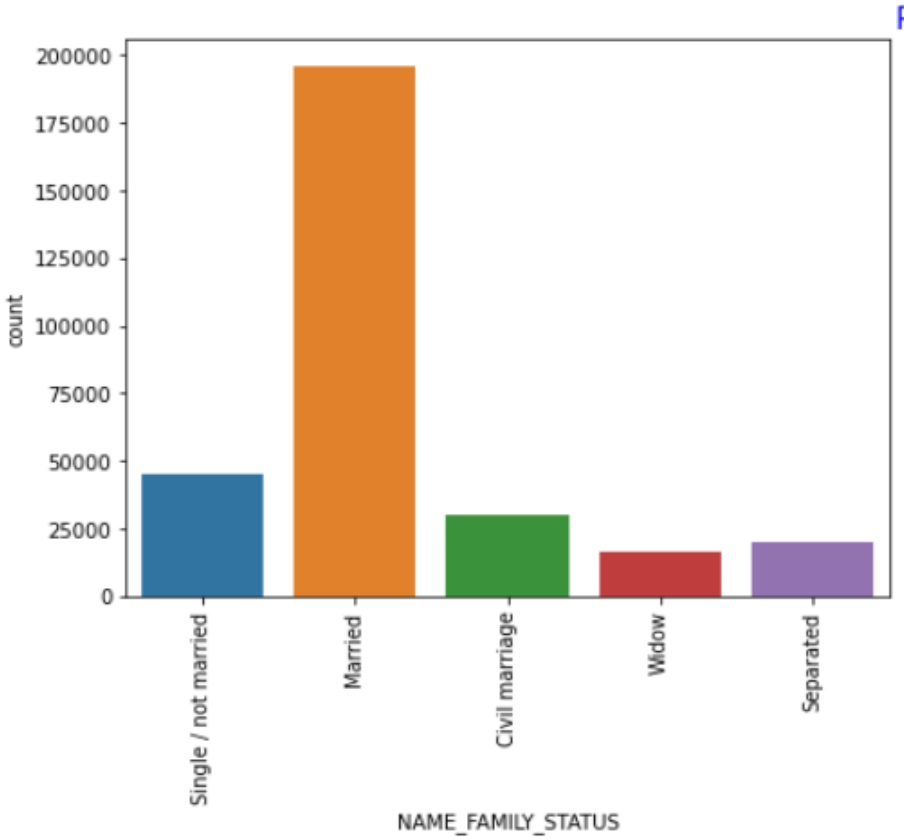


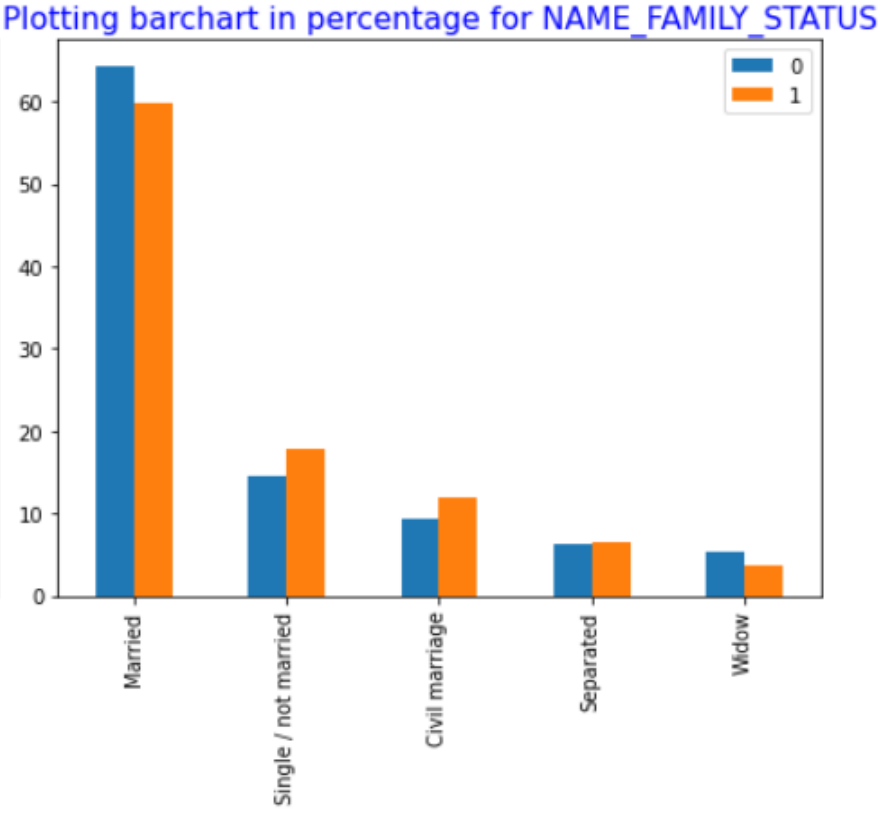
Observation:

- 1.People with secondary special and lower secondary education have more percentage of not returning loan on time.
- 2.People with higher education:98% and academic_degree:94% pay off the loan on time as compared to others.
- Hence, accepting loans of higher education people are less likely to become defaulters.

NAME_FAMILY_STATUS



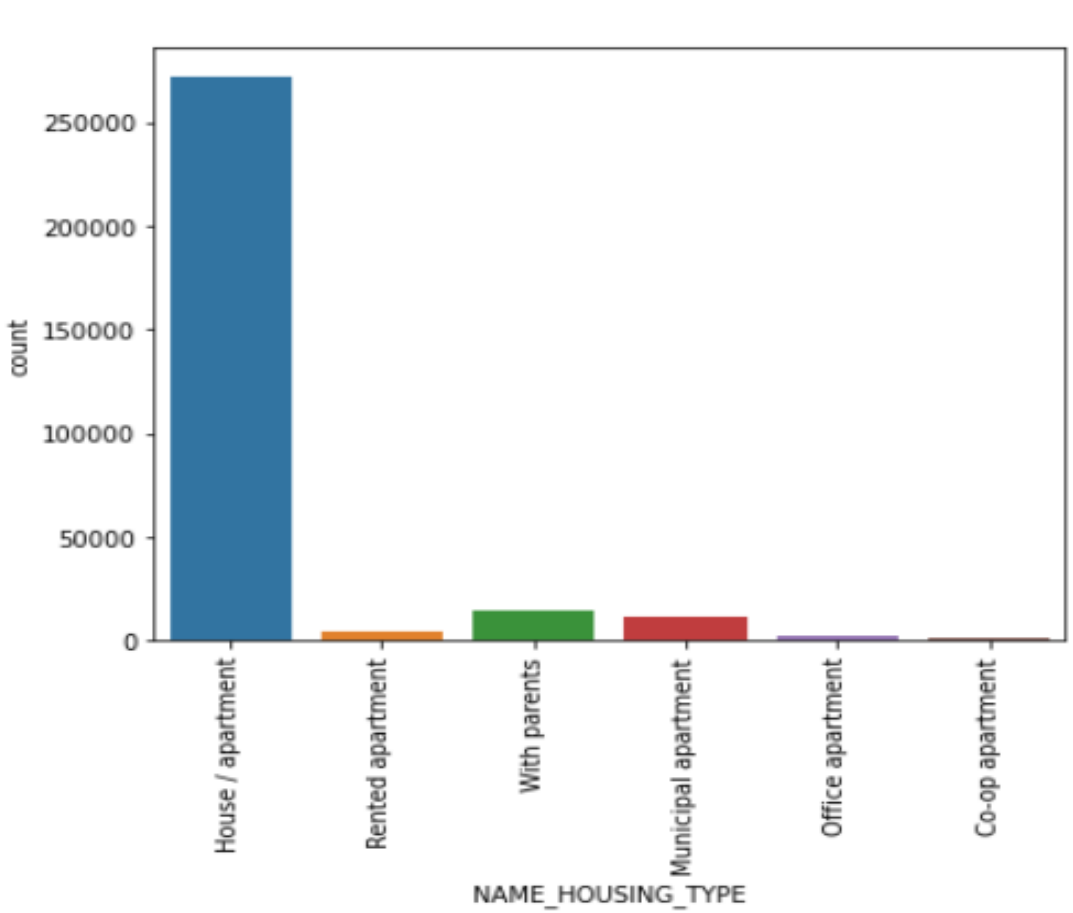


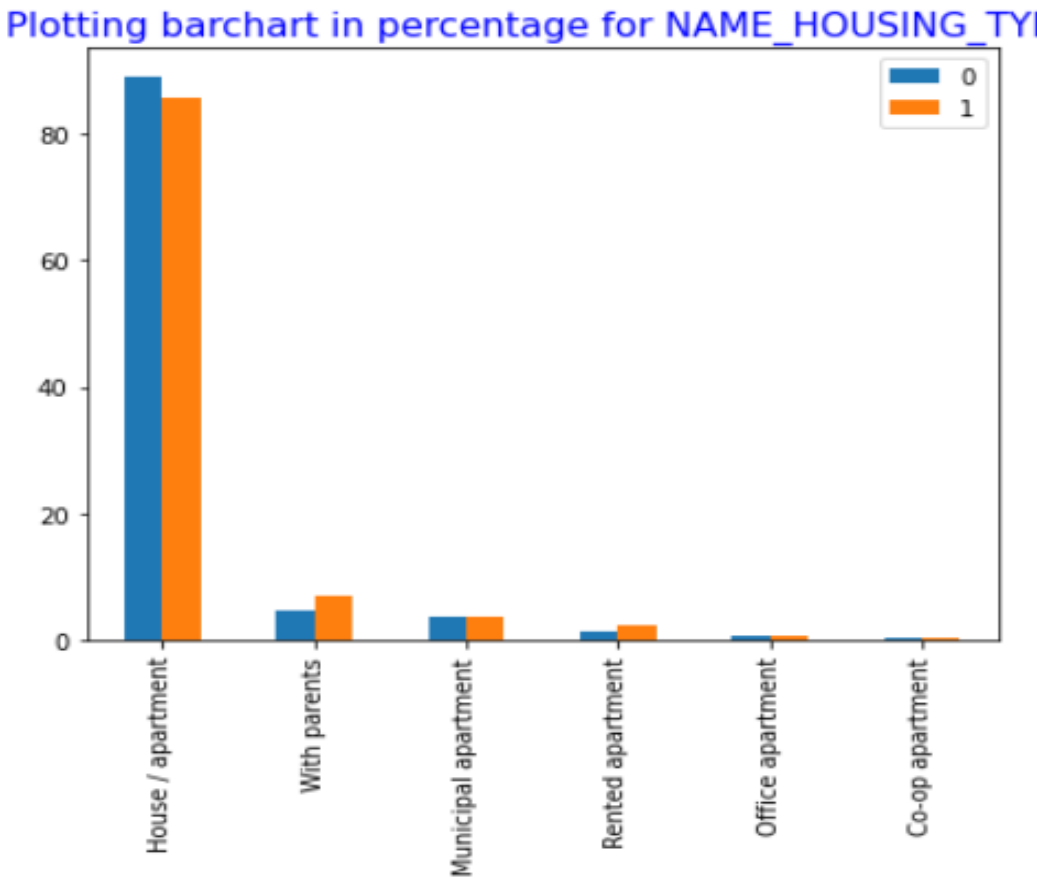


Observation:

- 1.Married people have highest counting in taking a loan followed by Single/not married.
- 2.Single/not married: 9.8%,civil marriage: 9.9% and separated people: 8.1% are more likely to become defaulter as compared to others
- 3.Married people: 92% can be preferred to give loan and widow: 92% too.

NAME_HOUSING_TYPE

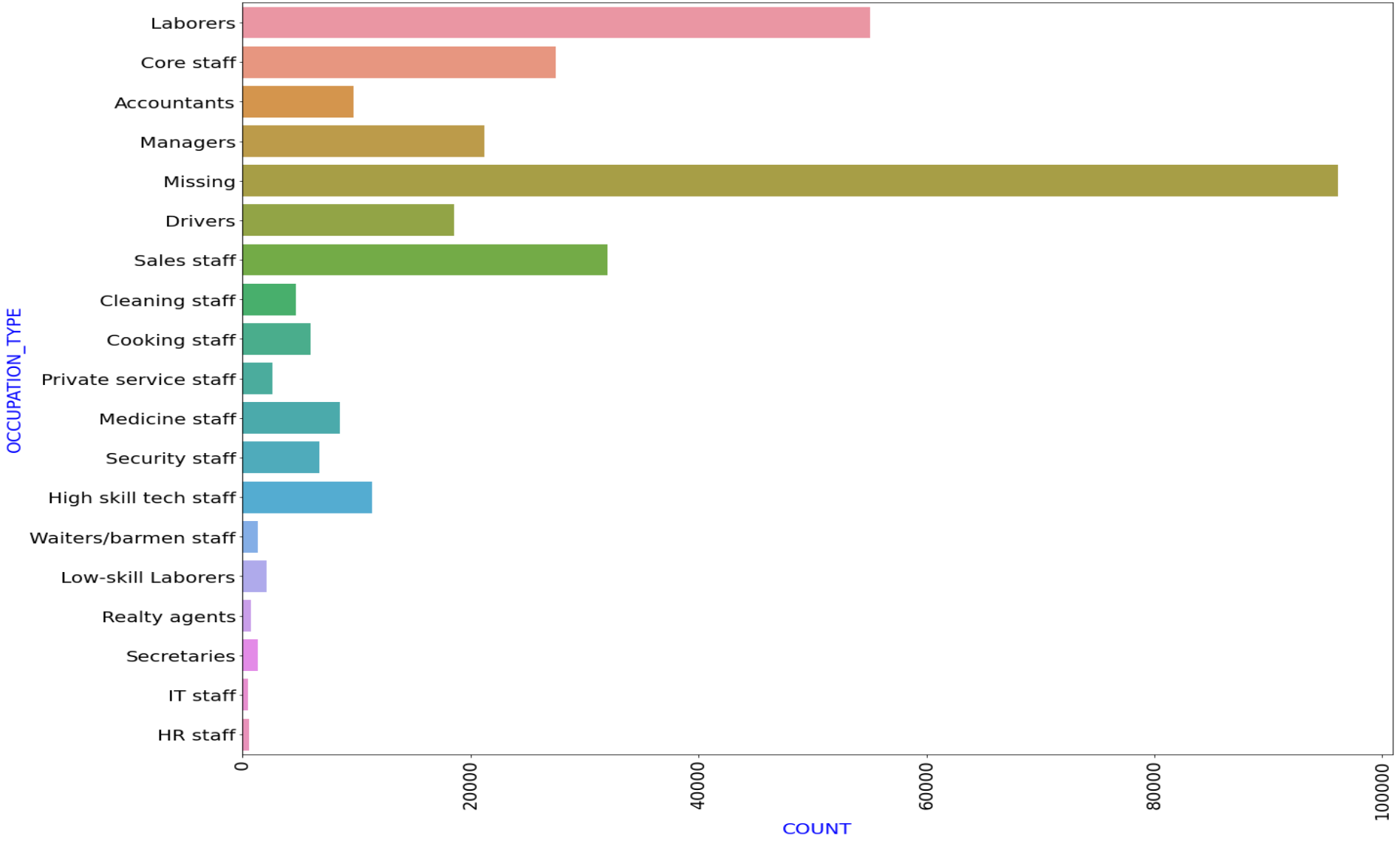


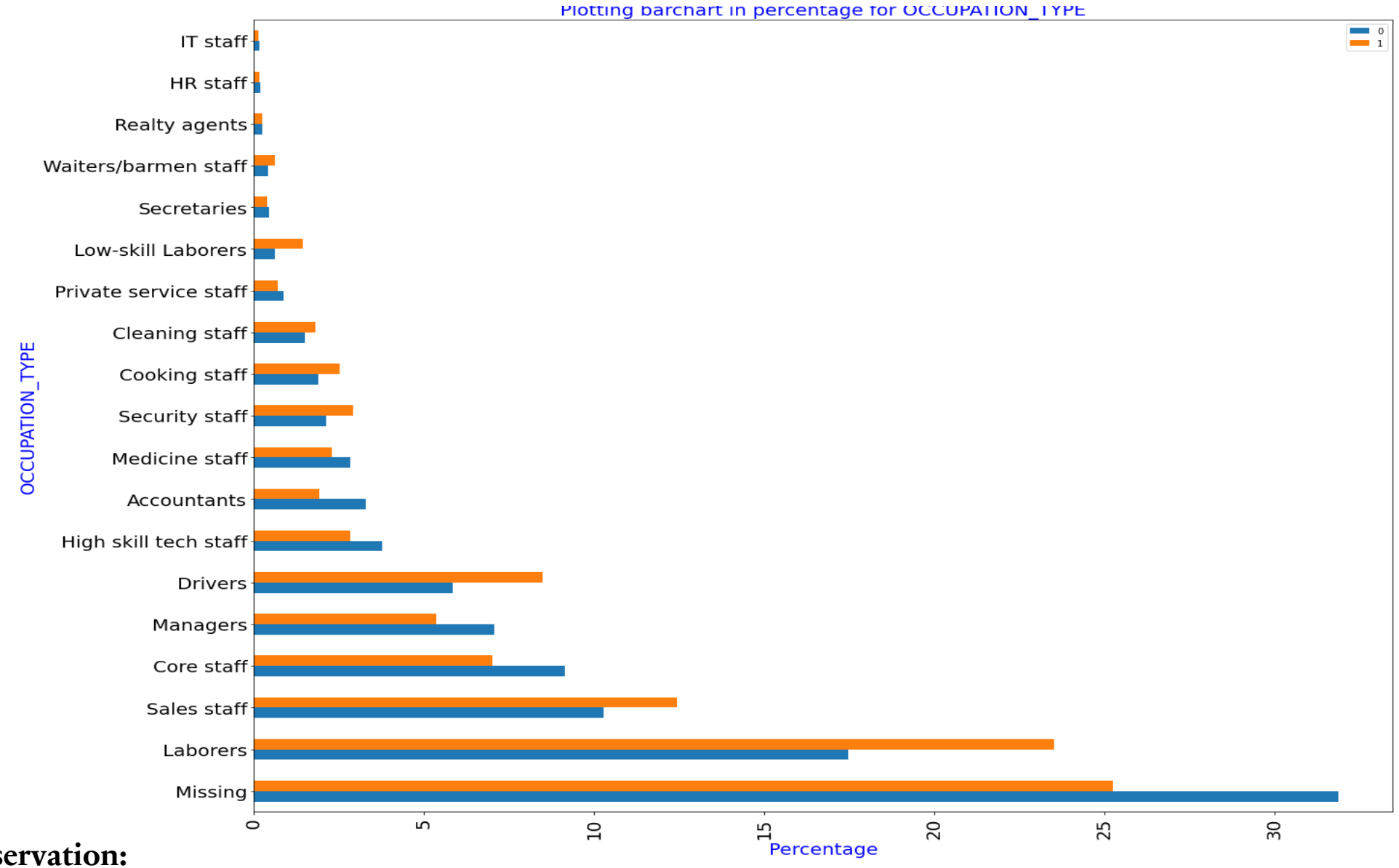


Observation:

- 1.People living in House/apartment: 92% have highest number in taking loan from the bank and pay on time.
- 2.People living with parents: 11% due to responsibility and rented apartment: 12% face difficulty to pay off the loan on time.
- 3.People living in House/apartment: 92% and office apartment: 93% are less likely to become defaulter as compared to others.

OCCUPATION_TYPE

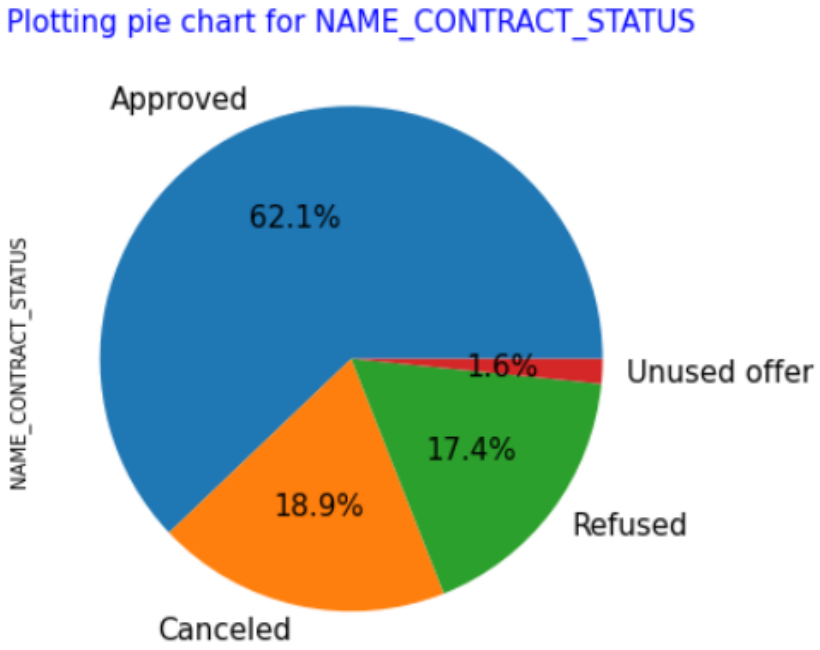


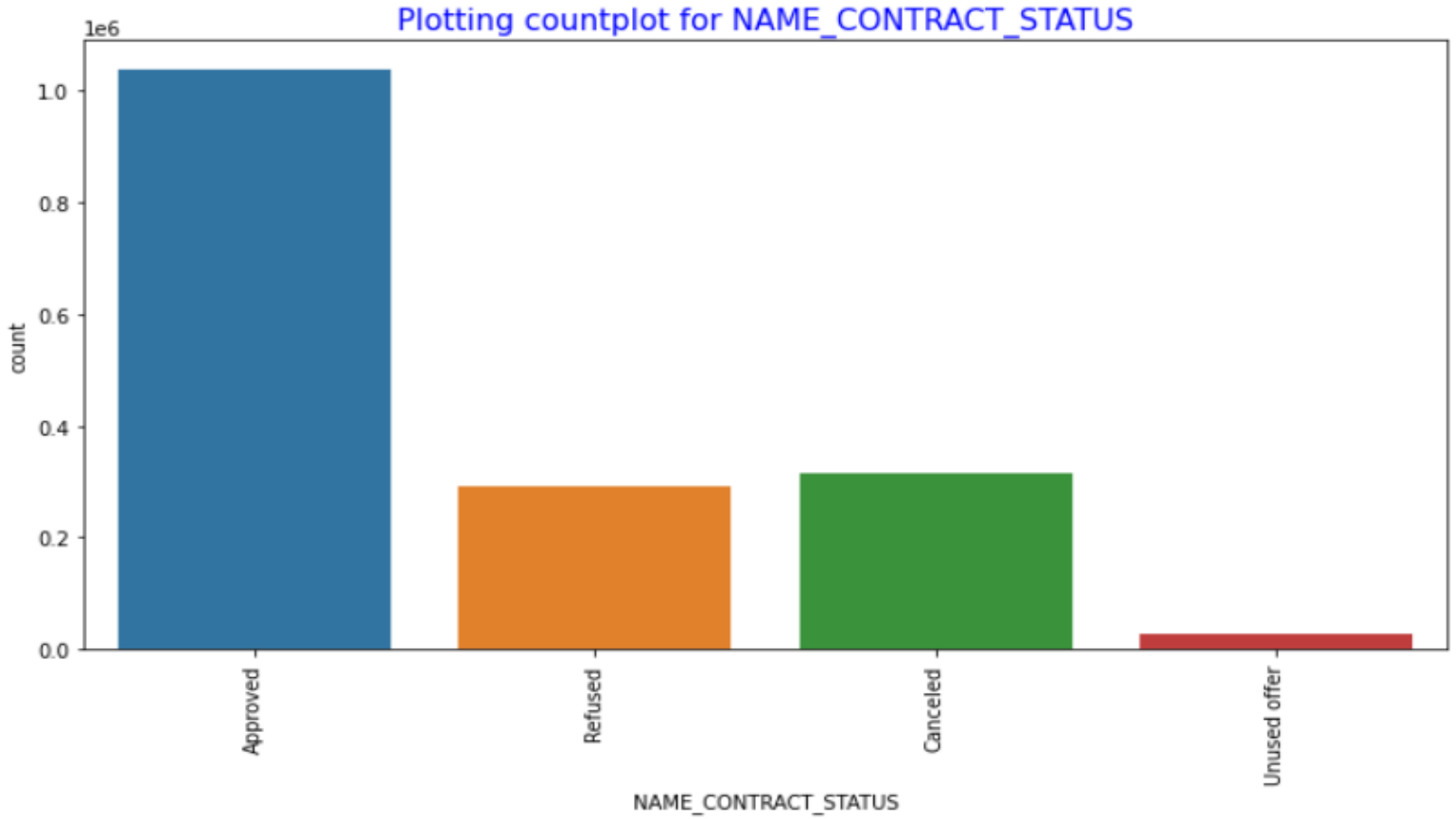


Observation:

- 1.laborers,sales staff and core_staff apply the most to take a loan from the bank.
- 2.Laborers, Sales staff, drivers and security staff fail to pay the loan on time.
- 3.Whereas core staff , High skill tech staff, Managers and Accountants are more likely to pay on time. It seems that people with good occupation earns more and hence pay off the loan on time.

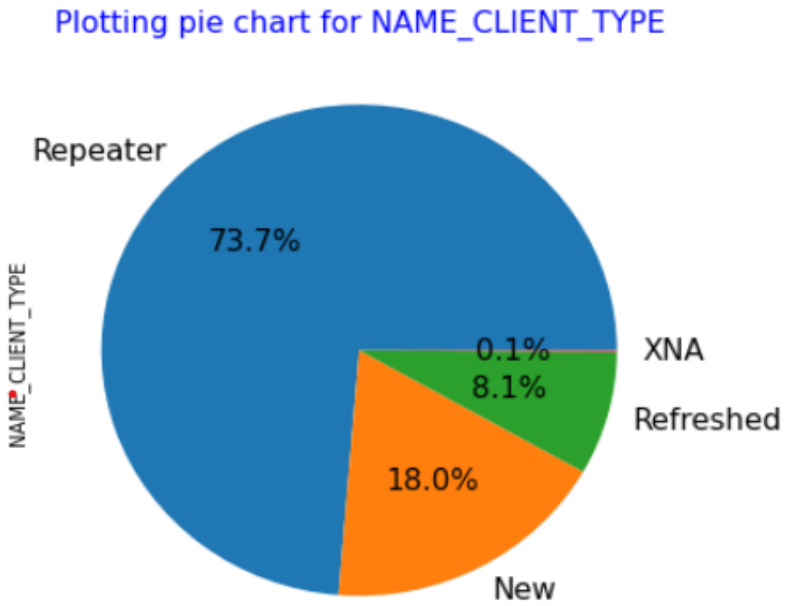
NAME_CONTRACT_STATUS

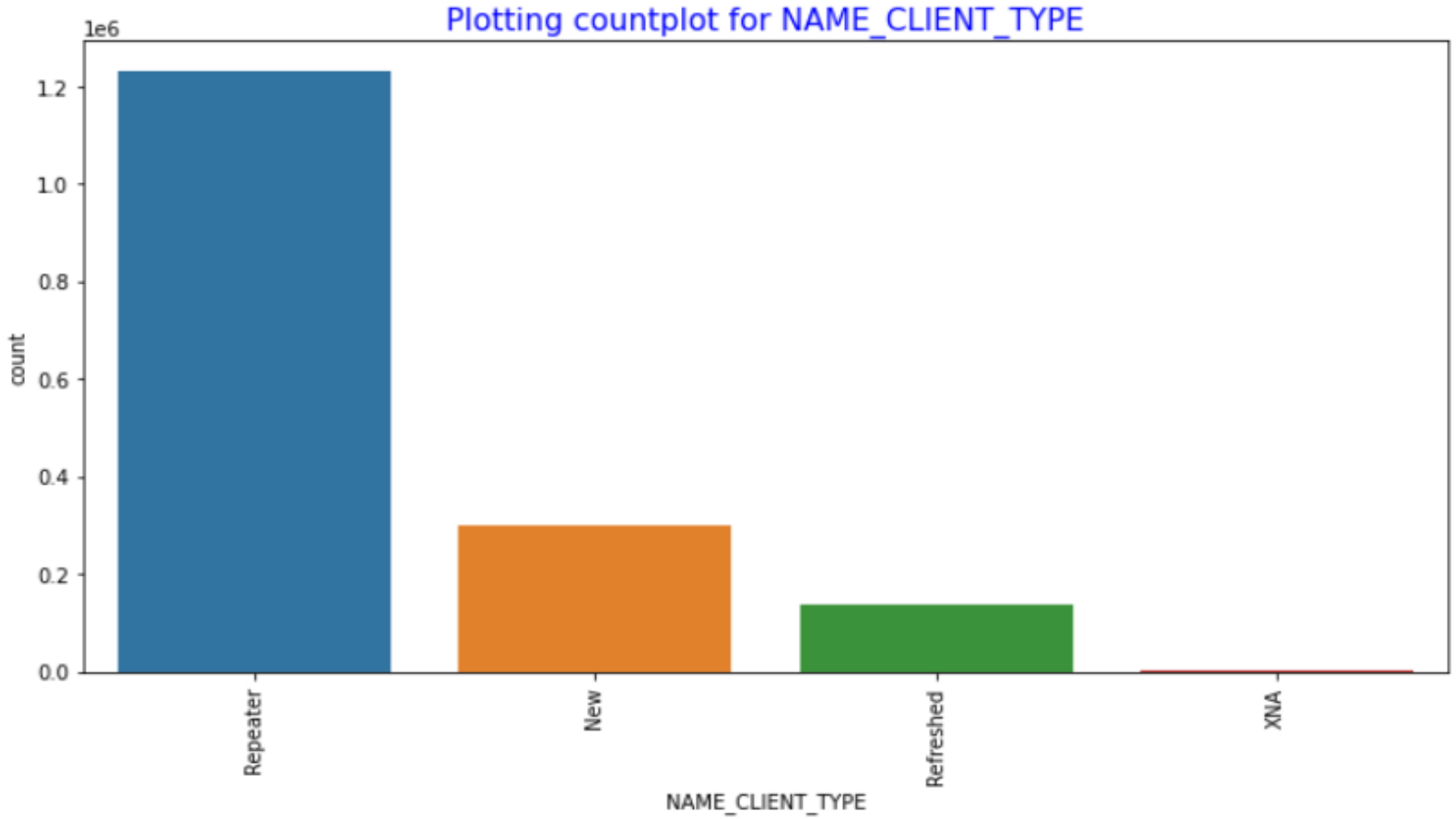




Observation: 62.1% of the loan application are being accepted,18.9% is cancelled and 17.4% is refused.

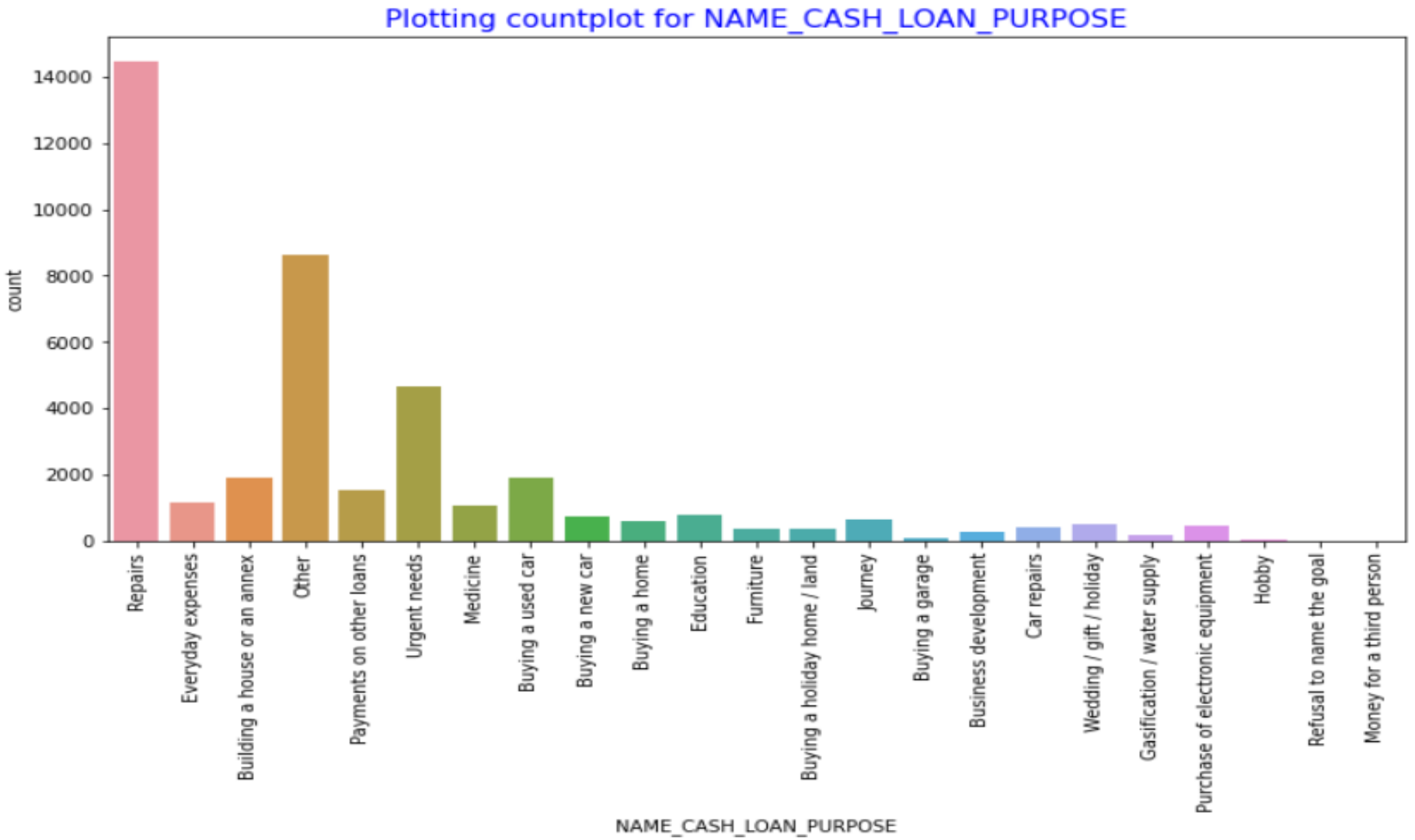
NAME_CLIENT_TYPE





Observation: 73.7% of people taking loan are repeater and 18% are taking loan for the first time.

NAME_CASH_LOAN_PURPOSE

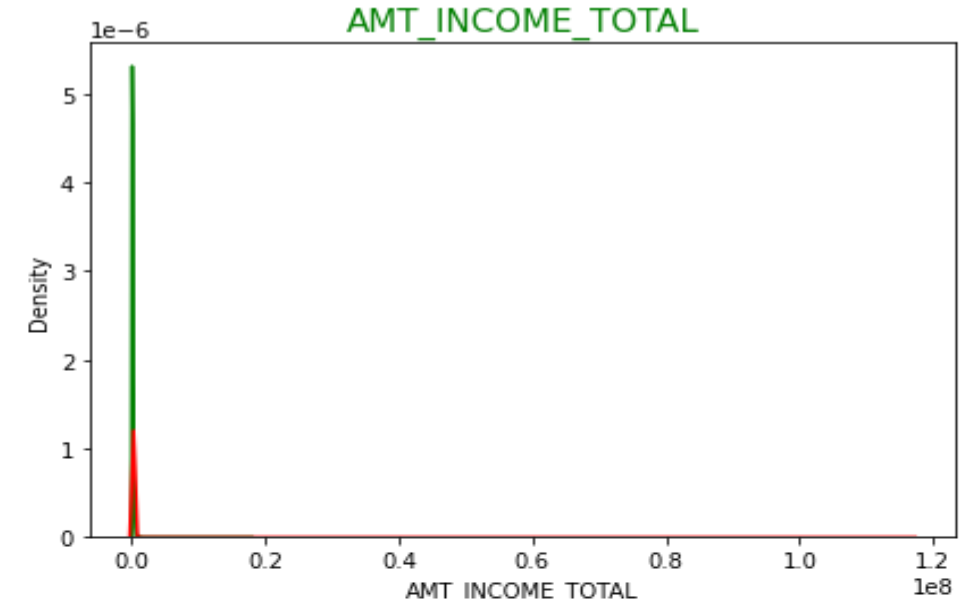


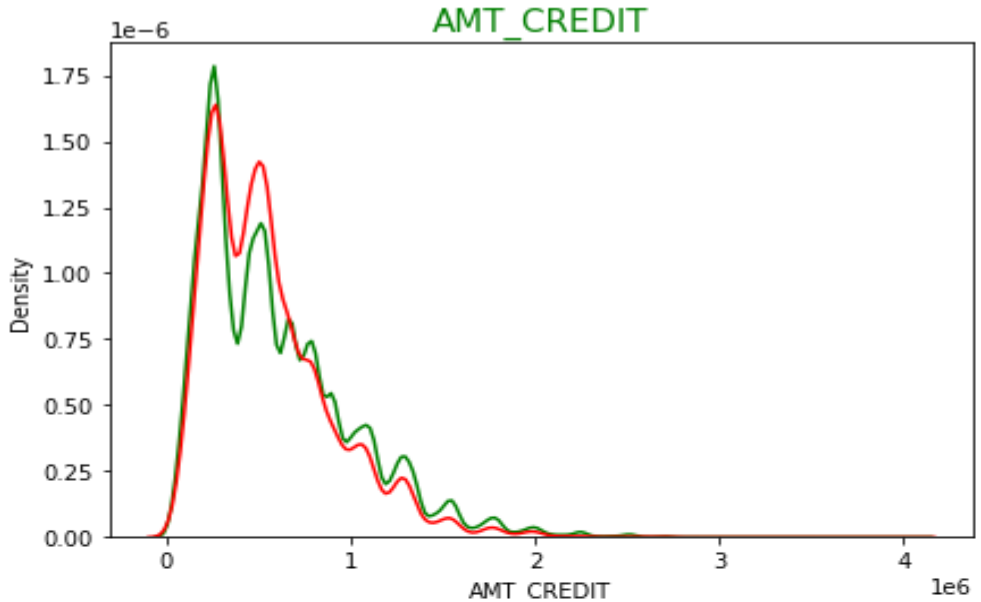
Observation: People take loan mostly for repairs, to build a house or annex and for their urgent needs.

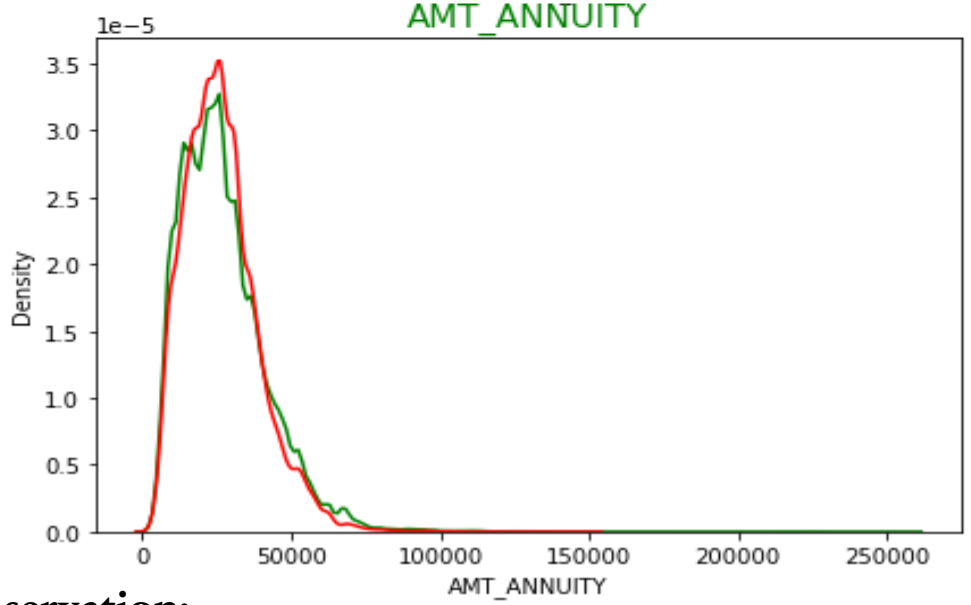
NUMERICAL UNIVARIATE ANALYSIS

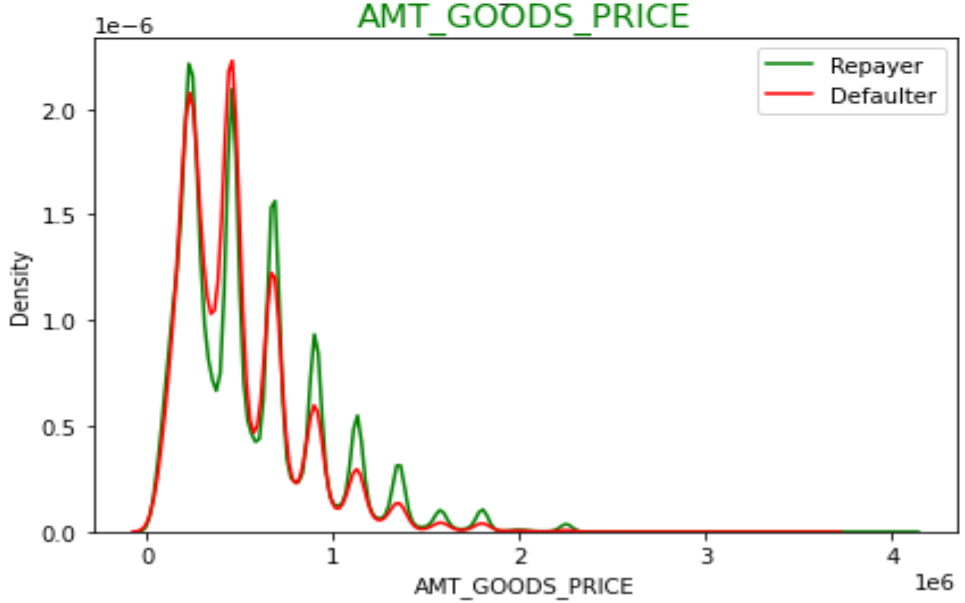
1. AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE
2. CLIENTS_AGE
3. CNT_CHILDREN
4. AMT_INCOME_TOTAL

AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE









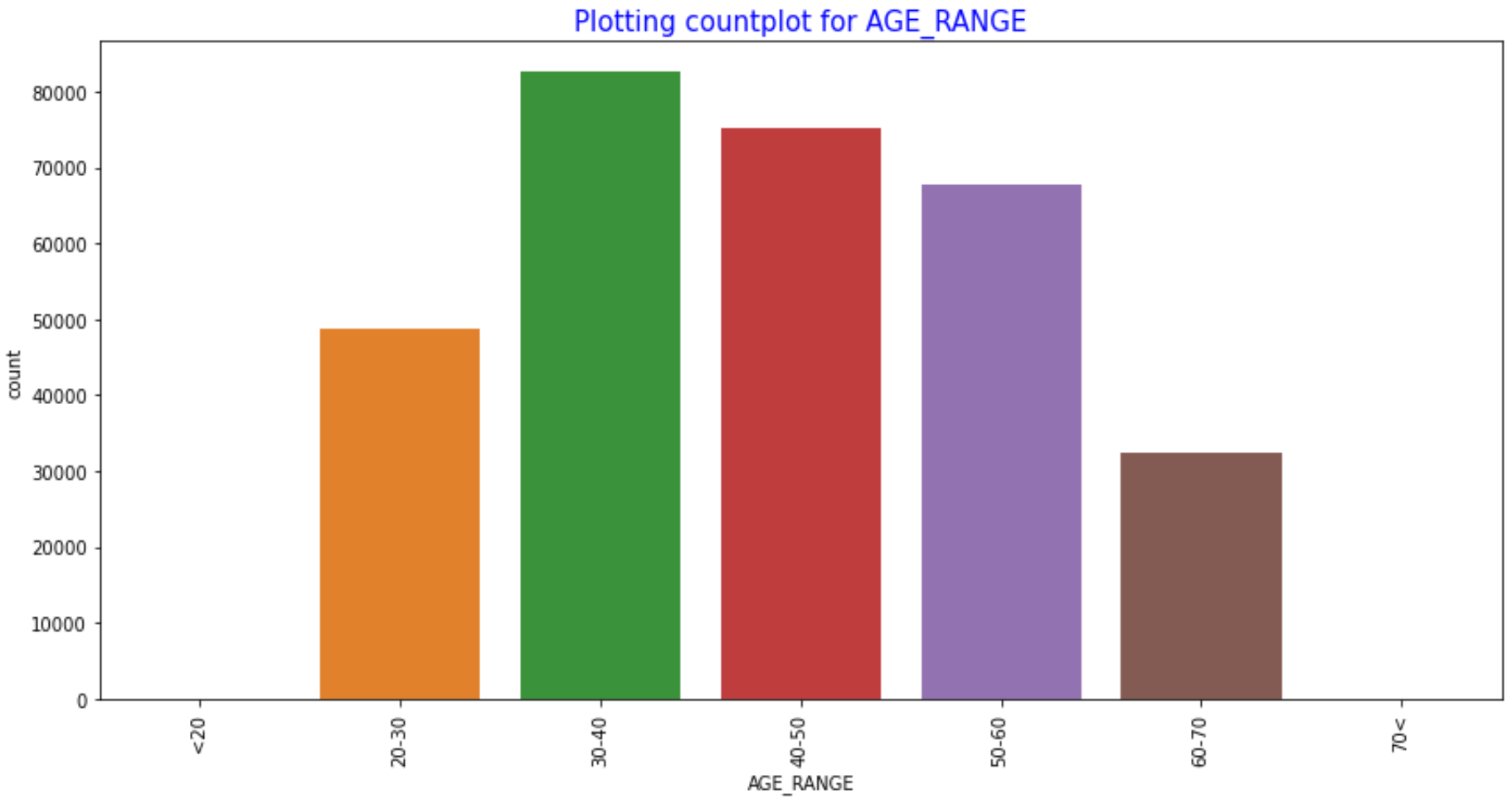
Observation:

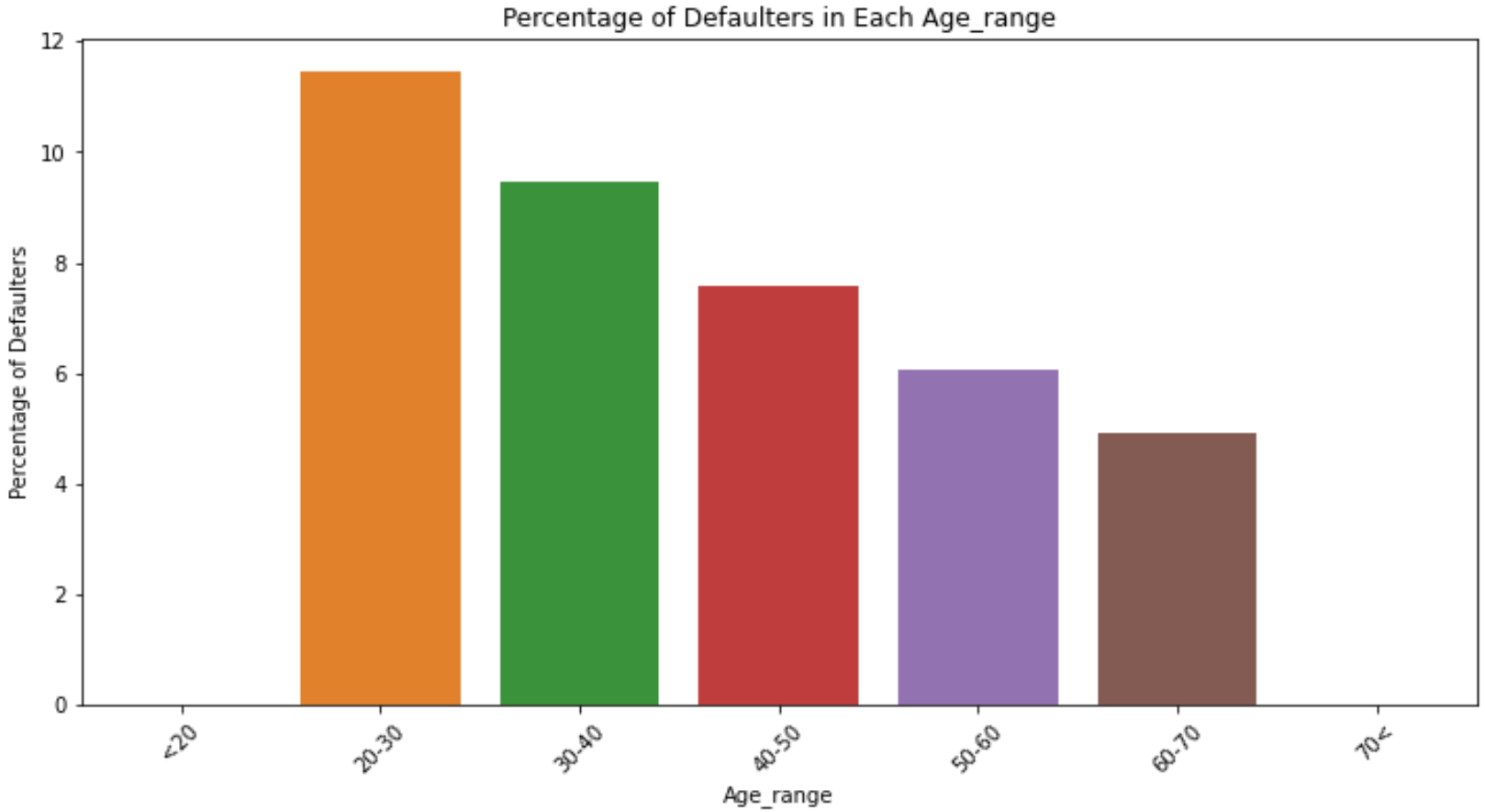
•1.AMT_CREDIT of loan is mostly less than 10,00,000. 2.Most people pay annuity below 50,000.

•3.Most people get loan for goods below 10,00,000

CLIENTS_AGE

30-40	0.269245
40-50	0.245544
50-60	0.220969
20-30	0.158806
60-70	0.105437
<20	0.000000
70<	0.000000



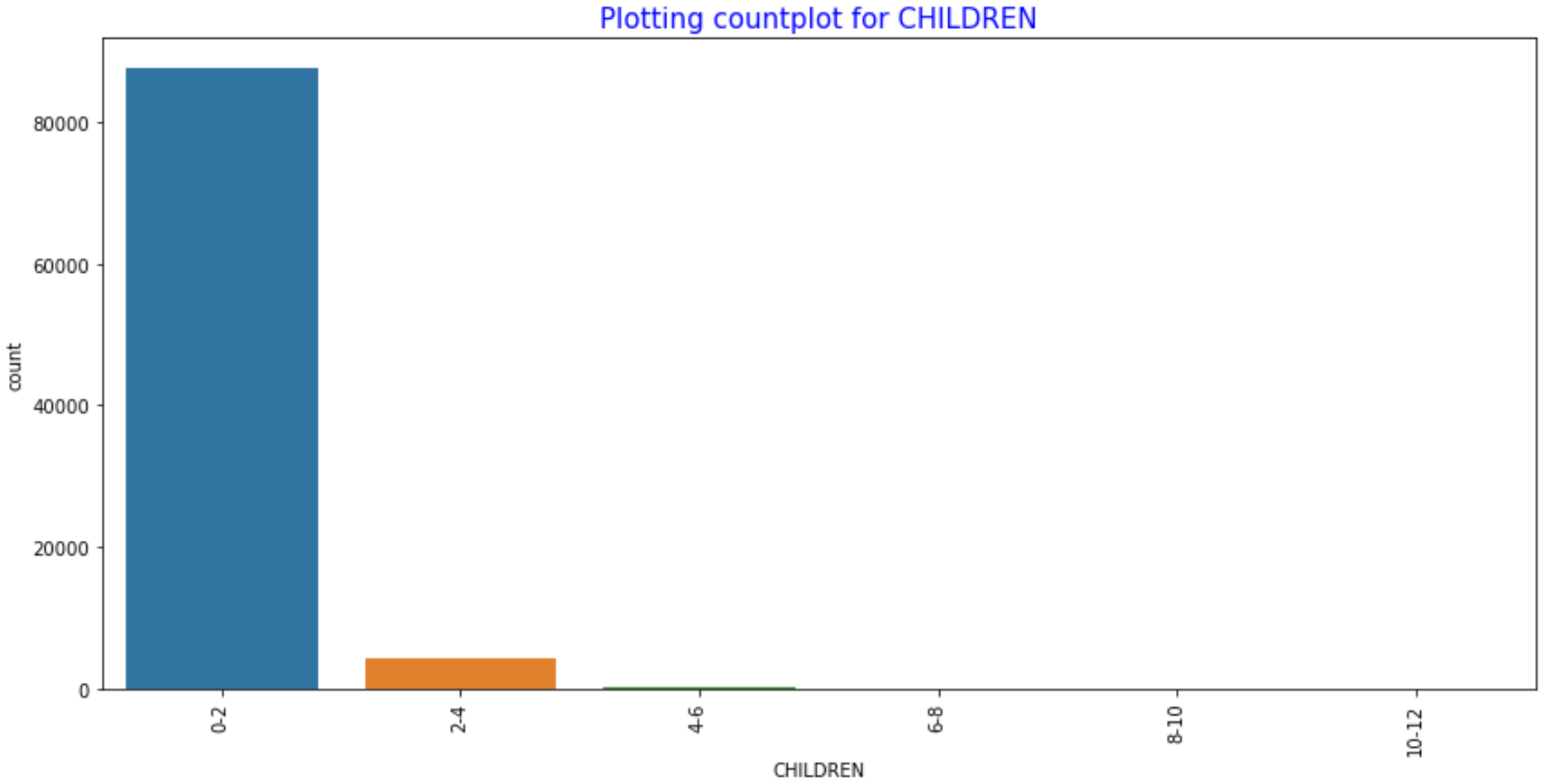


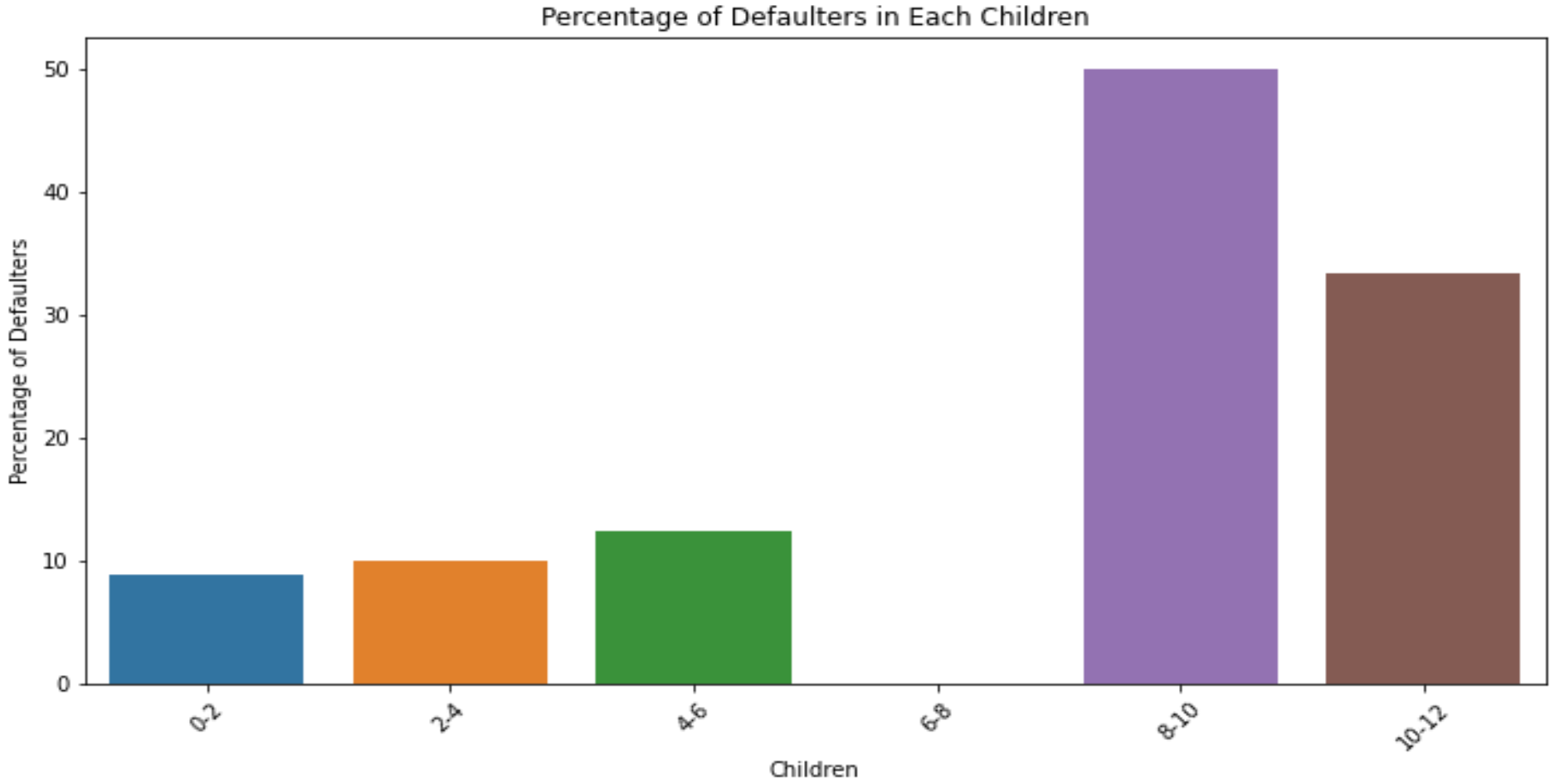
Observation:

- 1.Mostly people age between 30-40:26% are taking loans followed by 40-50:24% age group. 70+ age group people are less likely to take loans.
- 2.People in the range of 20-30 are more likely to become defaulters.
- 3.60-70 and more than 70 age group people are less likely to become defaulters.

CNT_CHILDREN

0-2	0.953641
2-4	0.045042
4-6	0.001143
6-8	0.000098
8-10	0.000044
10-12	0.000033



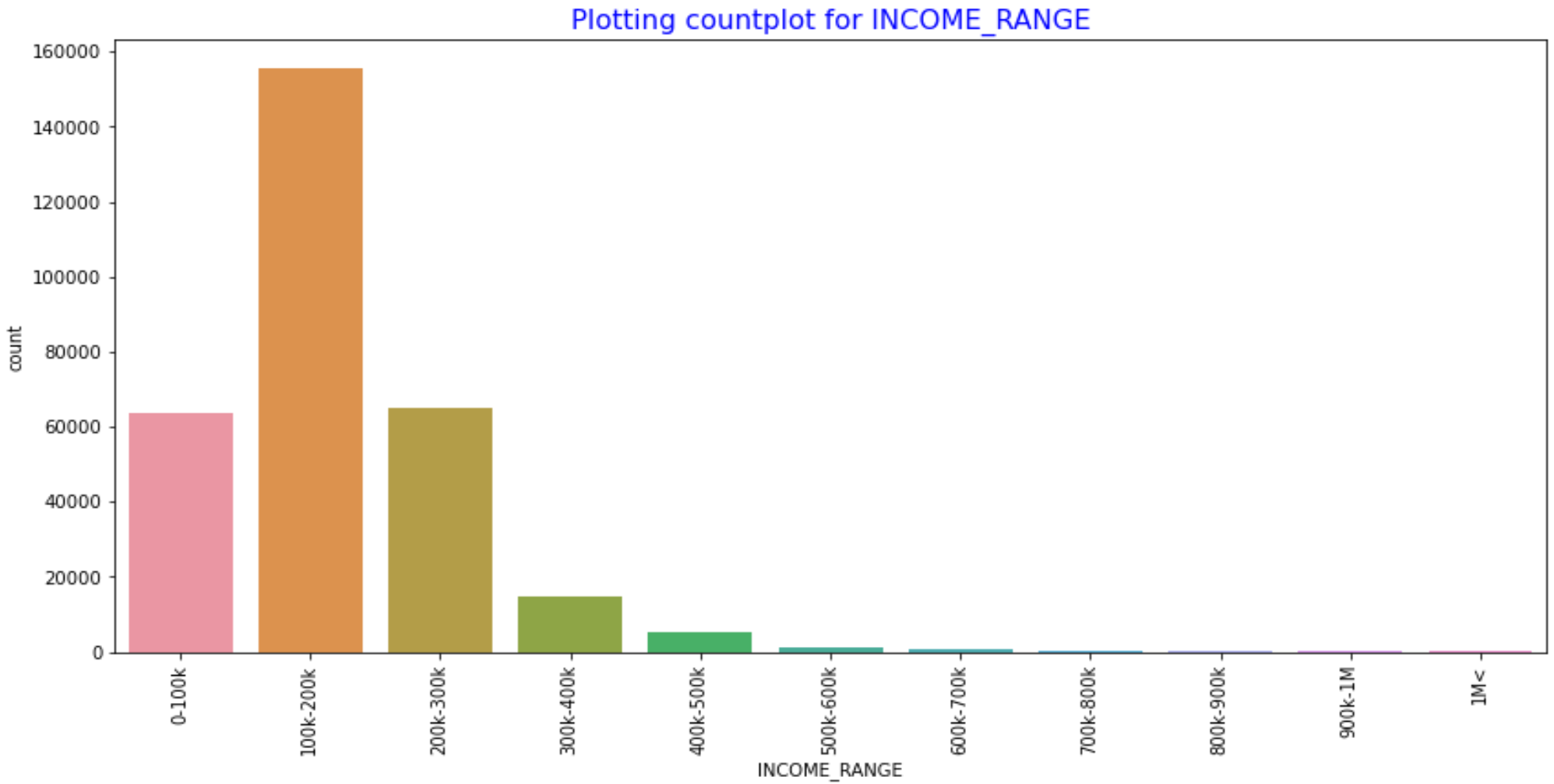


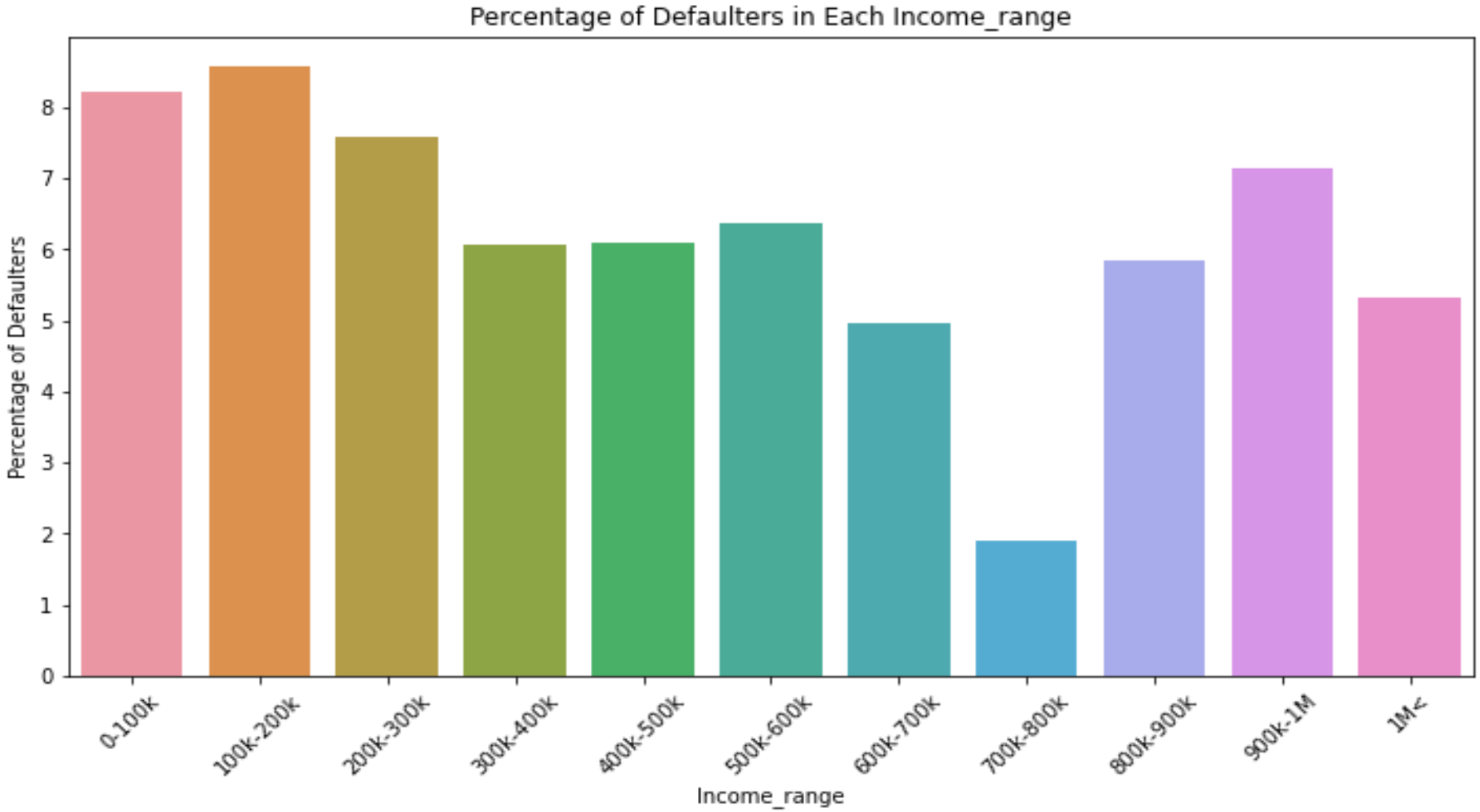
Observation:

- People with no children or at most 2 children which constitute 95% take more loans as compared to people with 6-8 children.
- People with more than 8 children are likely to become defaulters due to responsibilities.
- Hence, it is reliable to approve loan for people with 2 children or no children.

AMT_INCOME_TOTAL

100k-200k	50.728090
200k-300k	21.183933
0-100k	20.728449
300k-400k	4.765944
400k-500k	1.728621
500k-600k	0.353359
600k-700k	0.276684
800k-900k	0.094947
1M<	0.079612
700k-800k	0.051226
900k-1M	0.009136





Observation:

•1.More than 51% loan applicants have income between 100k-200k,91% people have income below 300k, people with income more than 700k are less likely to take out loan.

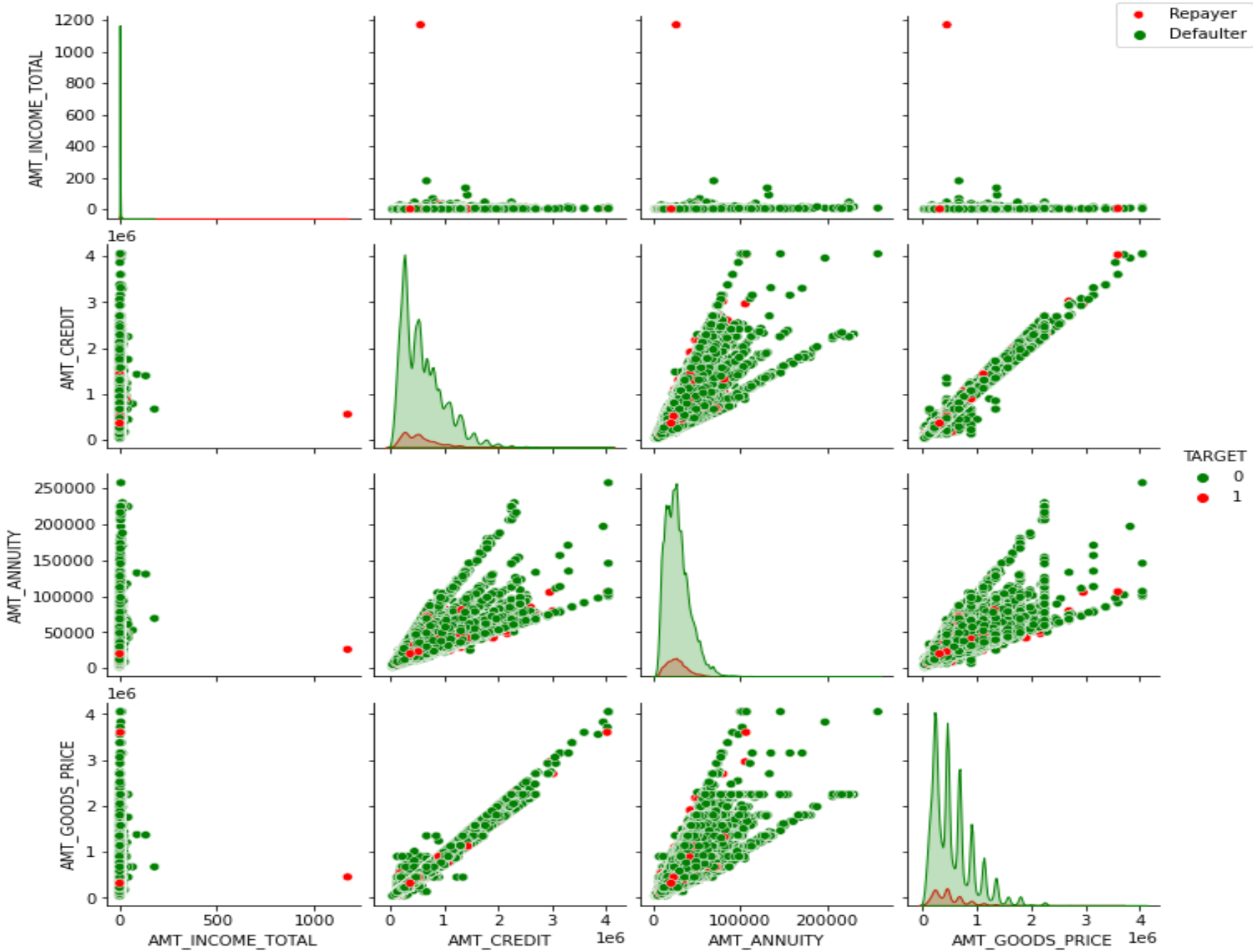
•2.People with income between 0-100k and 100k-200k become defaulter.

•3.Loan applicants with income between 700k-800k; 0.5% are less likely to become defaulter as compared to others.

BIVARIATE ANALYSIS

1. AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, TARGET
2. INCOME_RANGE VS NAME_CONTRACT_STATUS
3. AGE_RANGE VS NAME_FAMILY_STATUS
4. NAME_INCOME_TYPE VS CODE_GENDER
5. NAME_FAMILY_STATUS VS NAME_INCOME

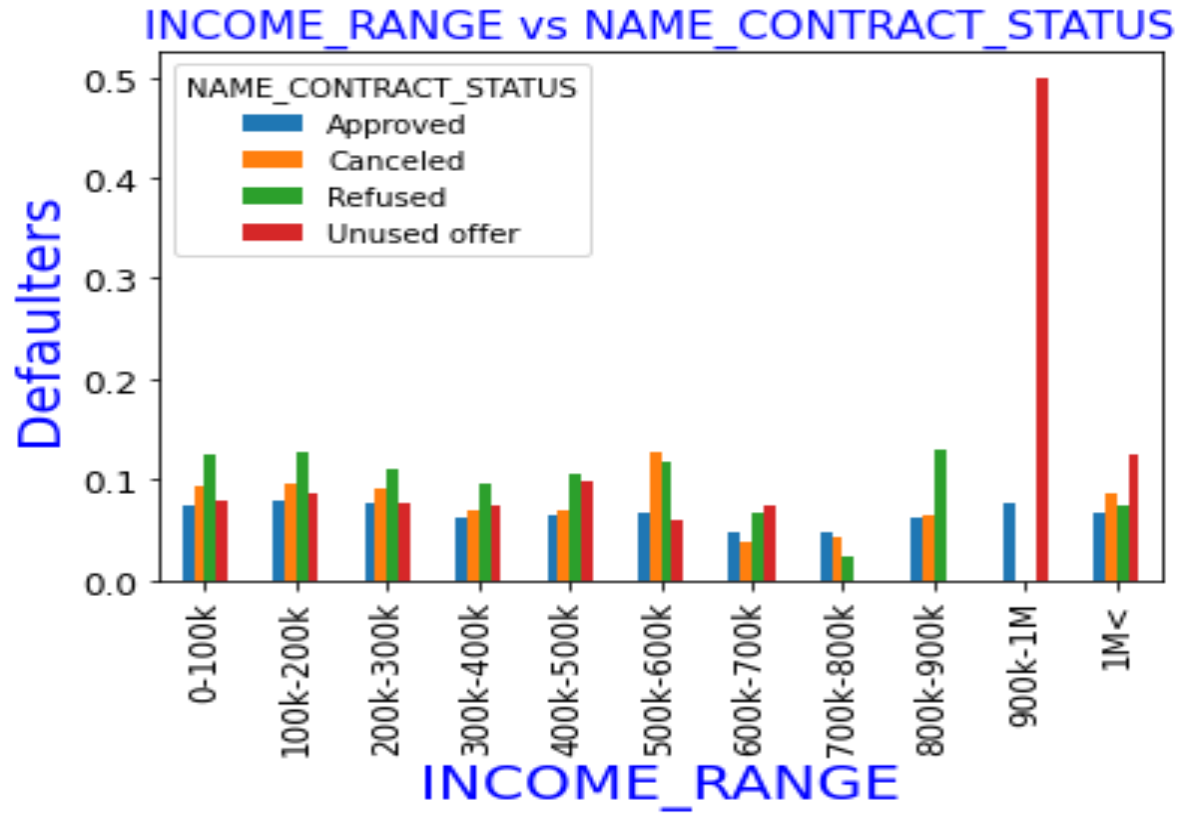
AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, TARGET



Observation: The correlation between AMT_CREDIT and AMT_GOODS_PRICE is high:0.99. People take credit more than 30,00,000 for goods prices and are likely to pay on time.

.

INCOME_RANGE VS NAME_CONTRACT_STATUS



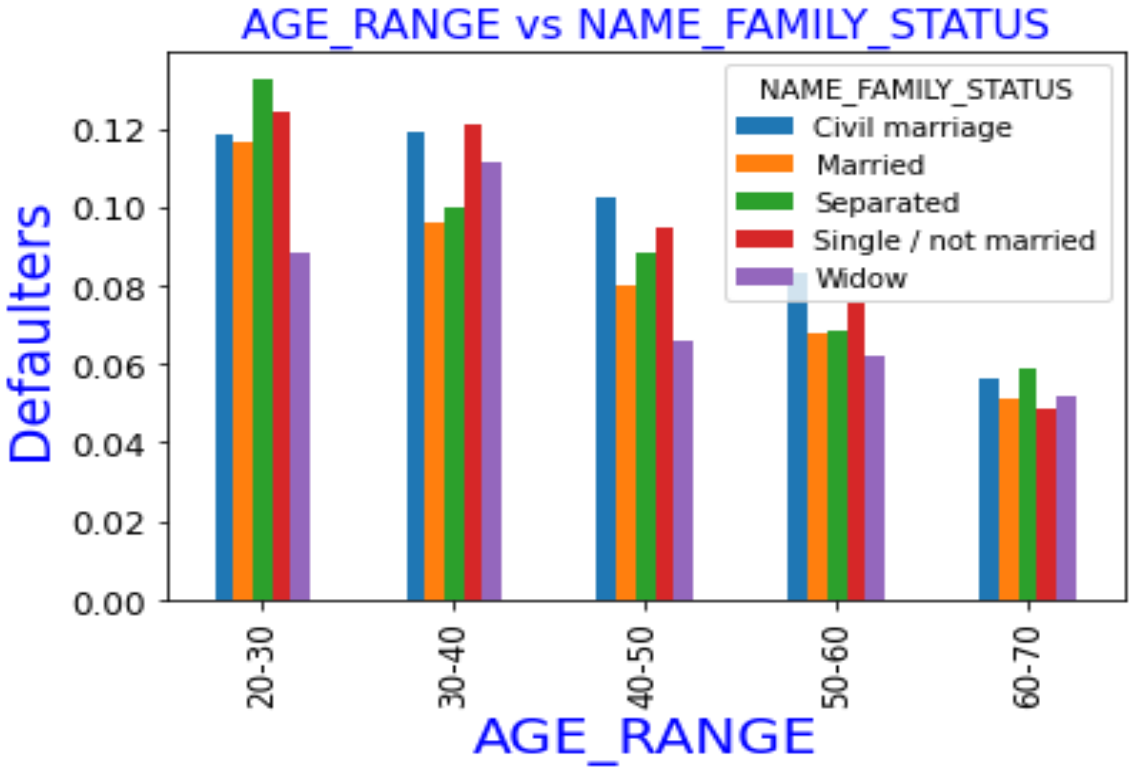
Observation:

•1.People with salary less than 300k have their previous application rejected and are more likely to become defaulters.

•2.People with salary 700k-800k have their previous application accepted and have less defaulters as compared to others.

•3.People with more than 1M income range have their previous loan status as unused offer.

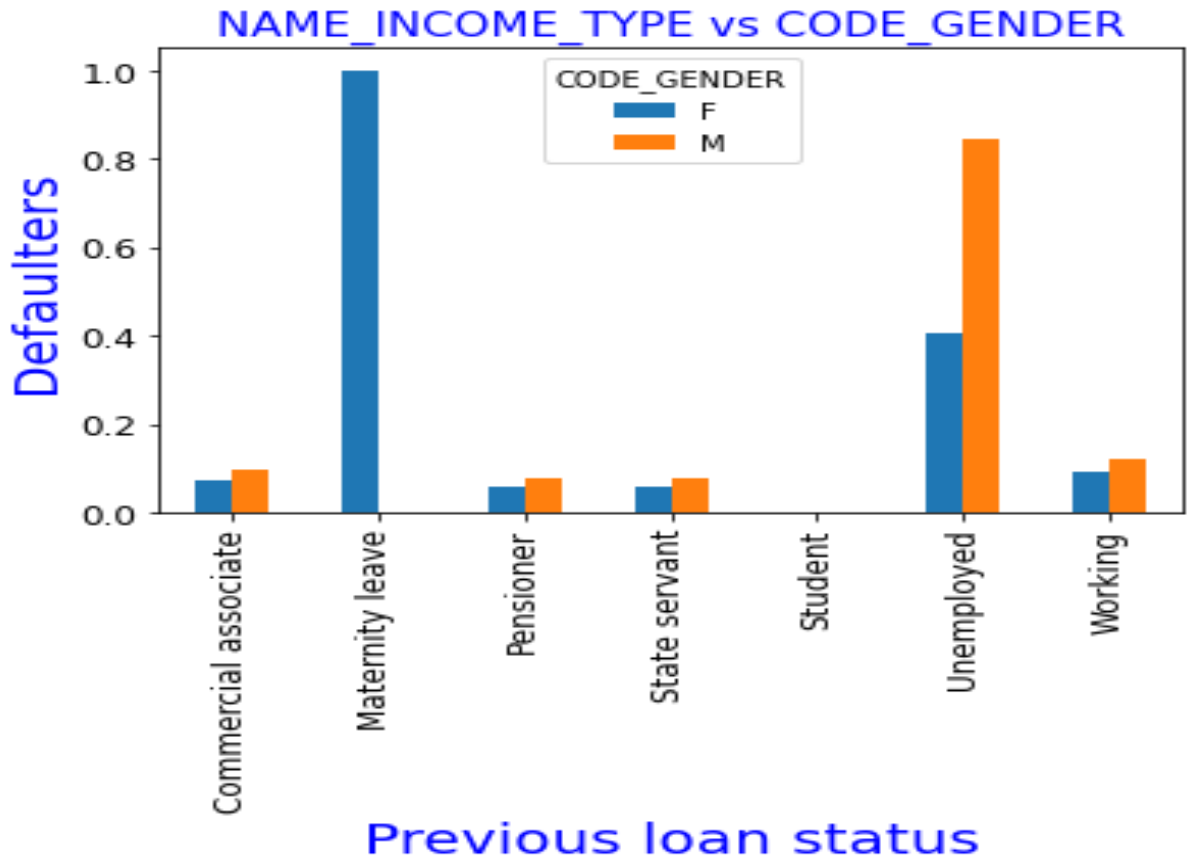
AGE_RANGE VS NAME_FAMILY_STATUS



Observation:

•1.Civil marriage in all age group are more likely to become defaulters.
•2.Widow between age group 30-40 fail to pay on time.
•3.Separated people between age group highly become defaulters.

NAME_INCOME_TYPE VS CODE_GENDER



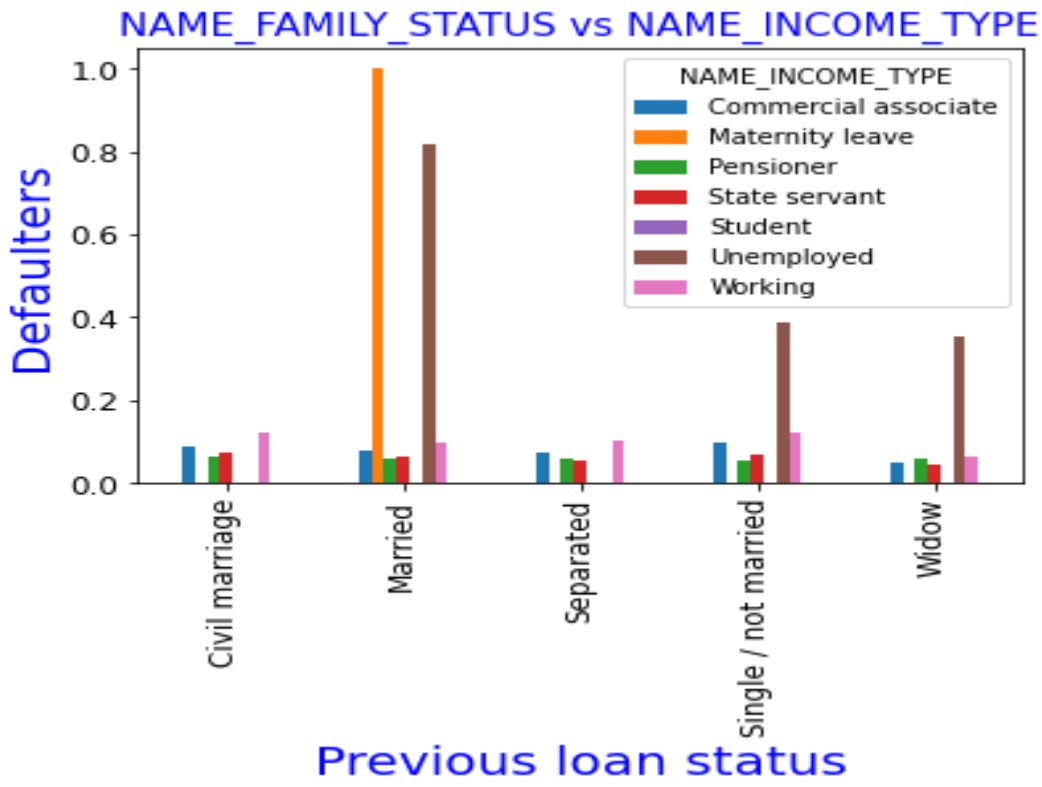
Observation:

•1.There are high chances for unemployed male to become a defaulter.

•2.Commercial associate, Pensioner and State servant are less likely to not pay on time.

•3.Female during maternity leave fails to pay on time and become defaulter.

NAME_FAMILY_STATUS VS NAME_INCOME

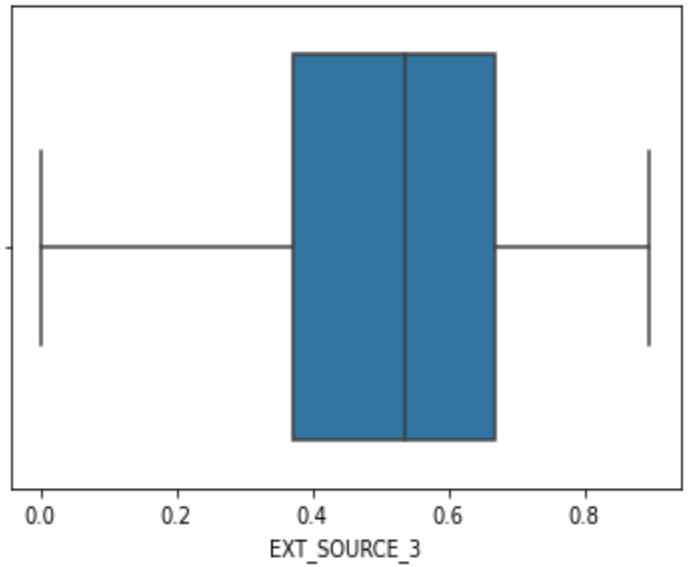


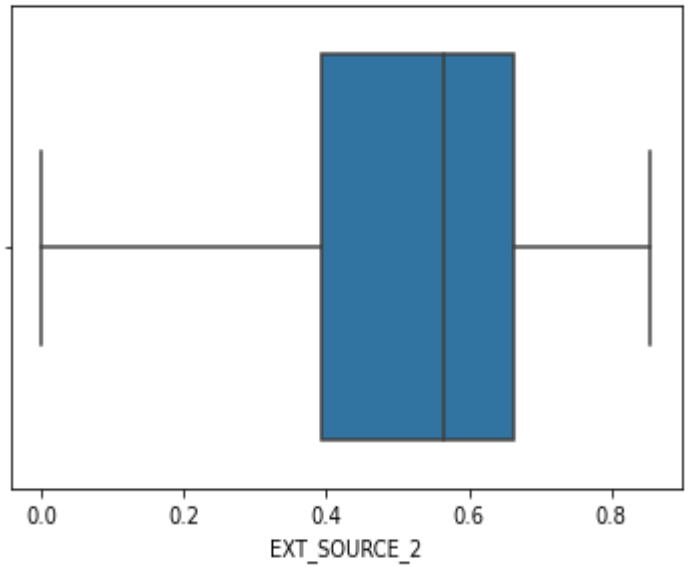
Observation:

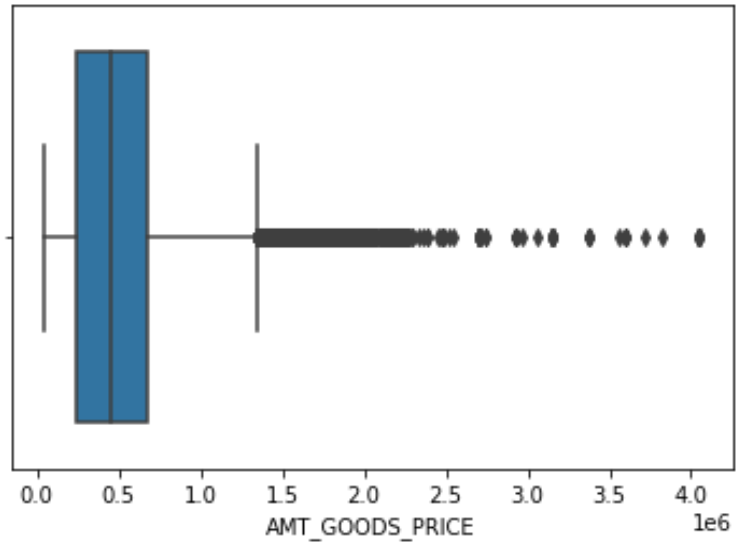
•1.Married -maternity leave, married-unemployed and Single-unemployed are more likely to fail to pay off the loan on time.

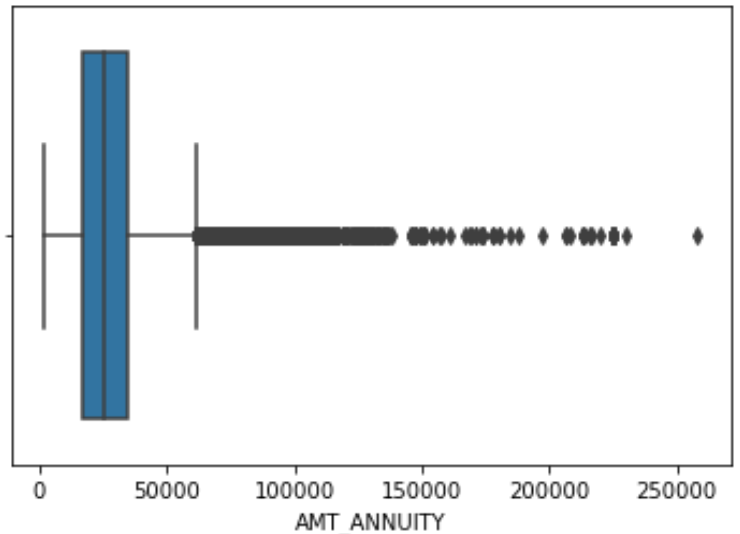
OUTLIERS

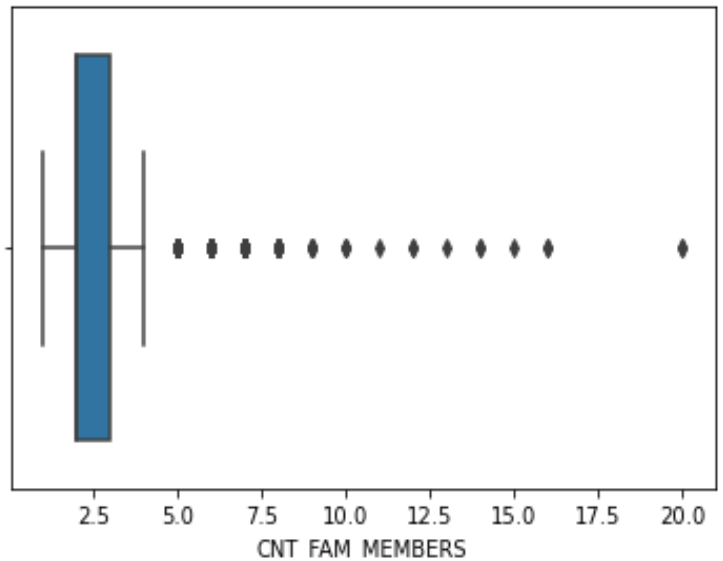
1. EXT_SOURCE_3
2. *EXT_SOURCE_2*
3. AMT_GOODS_PRICE
4. AMT_ANNUITY
5. CNT_FAM_MEMBERS


0.0
0.2
0.4
0.6
0.8
EXT_SOURCE_3


0.0
0.2
0.4
0.6
0.8
EXT_SOURCE_2


0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1e6
AMT_GOODS_PRICE


0
50000
100000
150000
200000
250000
AMT_ANNUITY


2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
CNT_FAM_MEMBERS

TOP 10 CORRELATION

Application Dataset

SK_ID_CURR	SK_ID_CURR	1.000000
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998508
AMT_CREDIT	AMT_GOODS_PRICE	0.987019
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950020
CNT_CHILDREN	CNT_FAM_MEMBERS	0.878640
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859332
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830429
AMT_ANNUITY	AMT_GOODS_PRICE	0.775548
AMT_CREDIT	AMT_ANNUITY	0.770515
DAYS_EMPLOYED	DAYS_BIRTH	0.626445

SK_ID_CURR	SK_ID_CURR	1.000000
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998269
AMT_CREDIT	AMT_GOODS_PRICE	0.982778
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.956655
CNT_CHILDREN	CNT_FAM_MEMBERS	0.885536
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.868994
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.778561
AMT_GOODS_PRICE	AMT_ANNUITY	0.752038
AMT_CREDIT	AMT_ANNUITY	0.751858
DAYS_BIRTH	DAYS_EMPLOYED	0.582706

Observation:

•1.The correlation is almost same for both repayer and defaulter.

•The logical correlation is between AMT_CREDIT and AMT_GOODS_PRICE ,CNT_CHILDREN and CNT_FAM_MEMBERS ,AMT_ANNUITY and AMT_GOODS_PRICE

Factors for an applicant to be a Defaulter, hence application can be approved.

1.CODE_GENDER	• The percentage of defaulter males : 10% is more. Hence, male can be a defaulter.
2. NAME_CONTRACT_TYPE	• More than 80% people taking cash loans fail to pay on time
3.NAME_INCOME_TYPE	• 40% of women with maternity leave become defaulters and unemployed : 42%.
4.NAME_EDUCATION_TYPE	• People with secondary special and lower secondary education have more percentage of not returning loan on time.
5.NAME_FAMILY_STATUS	• Single/not married: 9.8%,civil marriage: 9.9% and separated people: 8.1% are more likely to become defaulter.
6.OCCUPATION_TYPE	• Laborers, Sales staff, drivers and security staff fail to pay the loan on time.
7.CLIENTS_AGE	• People in the range of 20-30 are more likely to become defaulters.
8.CNT_CHILDREN	• People with more than 8 children are likely to become defaulters due to responsibilities.
9.AMT_INCOME_TOTAL	• People with income between 0-100k and 100k-200k become defaulter.

Factors for an applicant to be a potential repayer, hence application can be approved.

1.CODE_GENDER	• 65.8% females take loan. The number of females paying on time is almost double of number of men. Hence, accepting loans for females will be less risky than men.
2. NAME_CONTRACT_TYPE	• People taking revolving loans are likely to pay on time and the number of defaulter is also less. Hence, giving revolving loans will be less risky.
3.NAME_INCOME_TYPE	• Commercial associate ,Pensioner and State servant are more likely to pay on time. More than them student and businessman have no defaults.
4.NAME_EDUCATION_TYPE	• People with higher education:98% and academic_degree:94% pay off the loan on time as compared to others. Hence, accepting loans of higher education people are less likely to become defaulters.
5.NAME_FAMILY_STATUS	• widow: 94% can be preferred to give loan.
6.OCCUPATION_TYPE	• Core staff , High skill tech staff, Managers and Accountants are more likely to pay on time. It seems that people with good occupation earns more and hence pay off the loan on time.
7.CLIENTS_AGE	• 3.60-70 age group people are less likely to become defaulters.
8.CNT_CHILDREN	• People with no children or 2 children have less percentage of defaulters. Hence, it is reliable to approve loan for people with 2 children or no children
9.AMT_INCOME_TOTAL	• Loan applicants with income between 700k-800k; 0.5% are less likely to become defaulter as compared to others.