CREDIT EDA CASE STUDY

Member 1

Name: Sagar Kumar

Email Id: cse.sagarkumar@gmail.com

Member 2

Name: Nidhi Jha
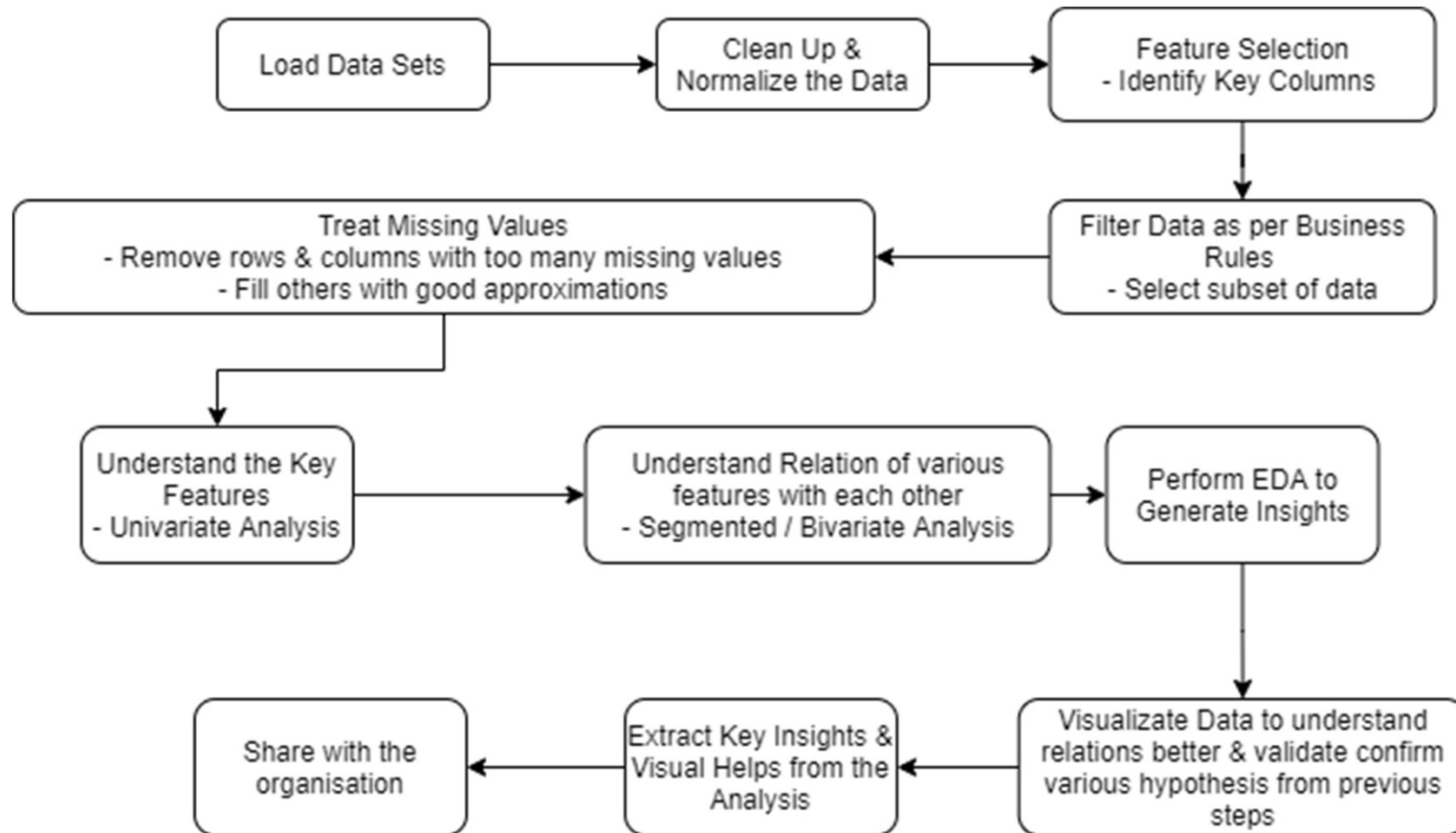
Email Id: nidjha669@gmail.com
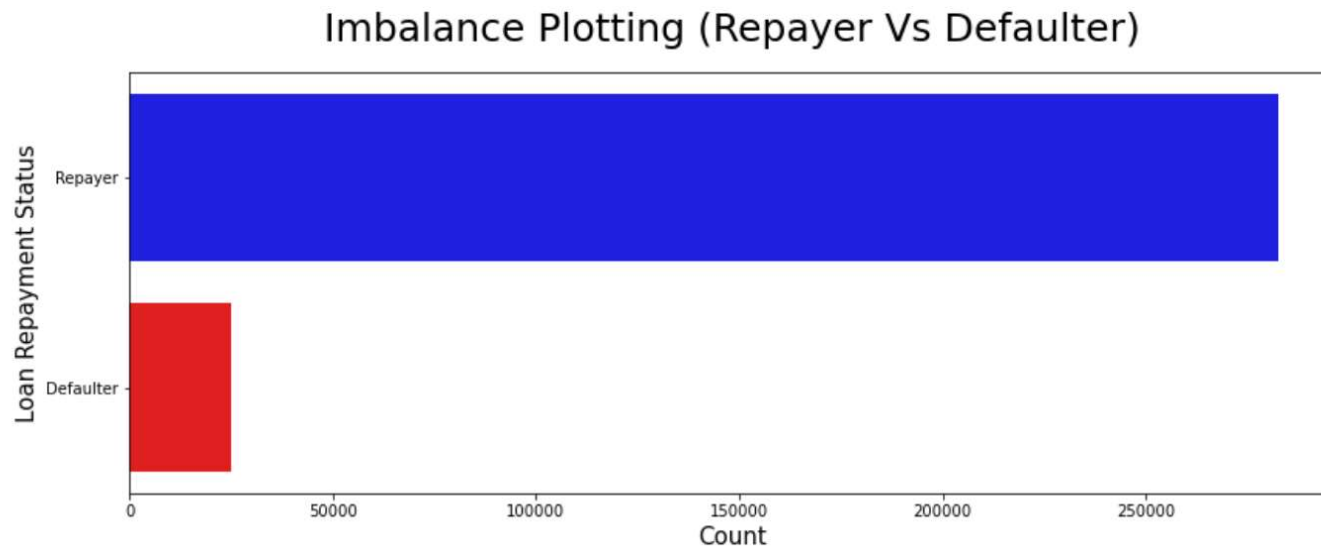
BUISNESS OBJECTIVE

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.

PROBLEM SOLVING METHODOLOGY

```
┌──────────────────┐      ┌──────────────────┐      ┌──────────────────┐
│  Load Data Sets  │ ───► │   Clean Up &     │ ───► │ Feature Selection│
│                  │      │ Normalize the Data│      │ - Identify Key   │
│                  │      │                  │      │   Columns        │
└──────────────────┘      └──────────────────┘      └──────────────────┘
                                                              │
                                                              ▼
┌────────────────────────────────────────────┐      ┌──────────────────┐
│         Treat Missing Values               │ ◄─── │ Filter Data as   │
│ - Remove rows & columns with too many      │      │ per Business     │
│   missing values                           │      │ Rules            │
│ - Fill others with good approximations     │      │ - Select subset  │
│                                            │      │   of data        │
└────────────────────────────────────────────┘      └──────────────────┘
        │
        ▼
┌──────────────────┐      ┌──────────────────────┐      ┌──────────────────┐
│ Understand the   │ ───► │ Understand Relation  │ ───► │ Perform EDA to   │
│ Key Features     │      │ of various features  │      │ Generate Insights│
│ - Univariate     │      │ with each other      │      │                  │
│   Analysis       │      │ - Segmented /        │      │                  │
│                  │      │   Bivariate Analysis │      │                  │
└──────────────────┘      └──────────────────────┘      └──────────────────┘
                                                                  │
                                                                  ▼
┌──────────────────┐      ┌──────────────────────┐      ┌──────────────────────┐
│  Share with the  │ ◄─── │ Extract Key Insights │ ◄─── │ Visualize Data to    │
│  organisation    │      │ & Visual Helps from  │      │ understand relations │
│                  │      │ the Analysis         │      │ better & validate    │
│                  │      │                      │      │ confirm various      │
│                  │      │                      │      │ hypothesis from      │
│                  │      │                      │      │ previous steps       │
└──────────────────┘      └──────────────────────┘      └──────────────────────┘
```

Analysis – Identifying Data Imbalance

Imbalance Plotting (Repayer Vs Defaulter)

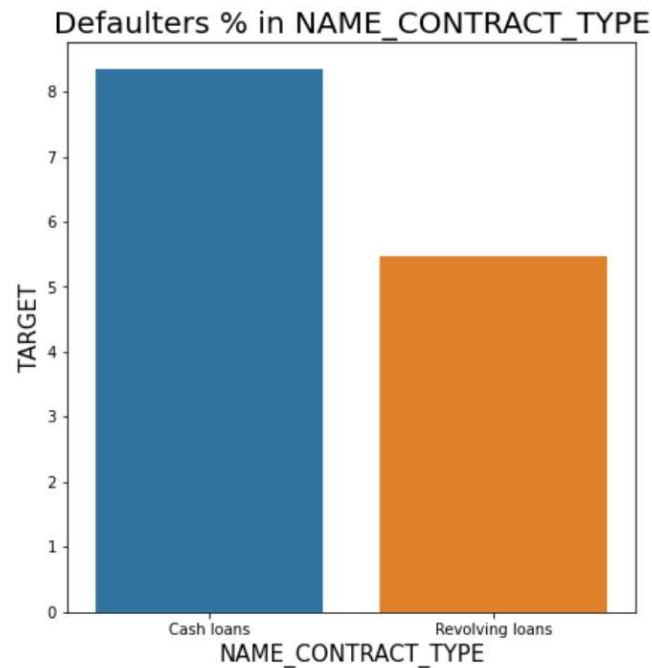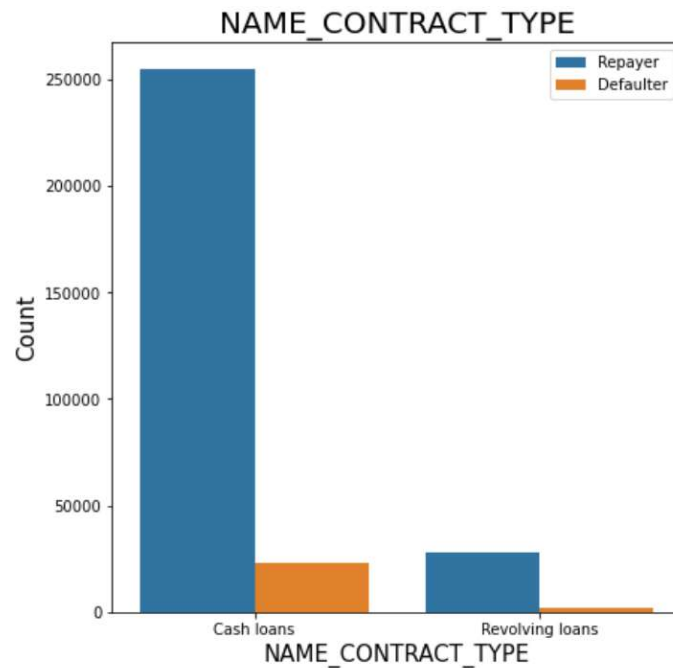Re-Payer Percentage is 91.93%.

Defaulter Percentage is 8.07%.

Imbalance Ratio: Re-Payer to Defaulter is: 11.39/1 (approx.)

UpGrad

1. Contract Type

Inferences:

- Revolving loans are just a small fraction (10%) from the total number of loans.
- Around 8-9% of Cash loans applicant and 5-6% Revolving loans applicant are defaulters.

2. Gender

- The number of female clients is slightly higher than the number of male clients.
- Based on the percentage of defaulted credits, males have a higher chance of not returning their loans about 10%, compared to women about 7%.
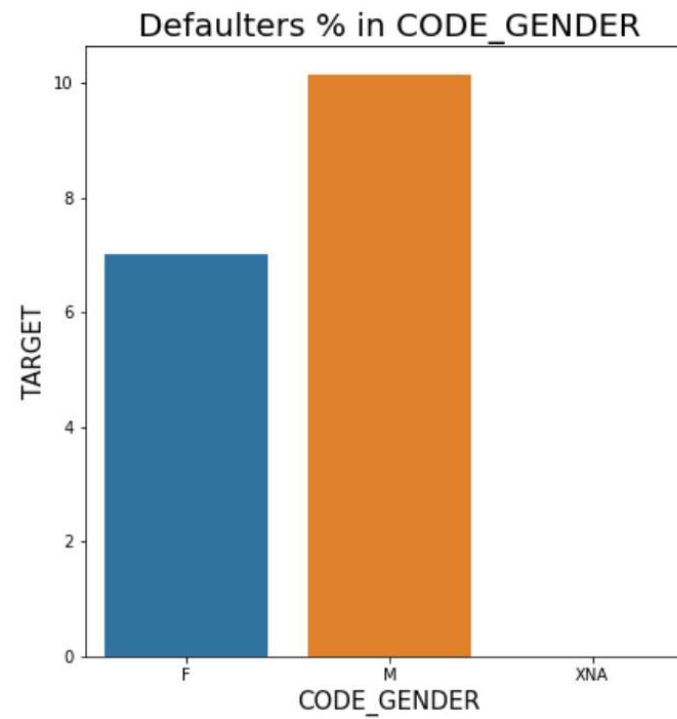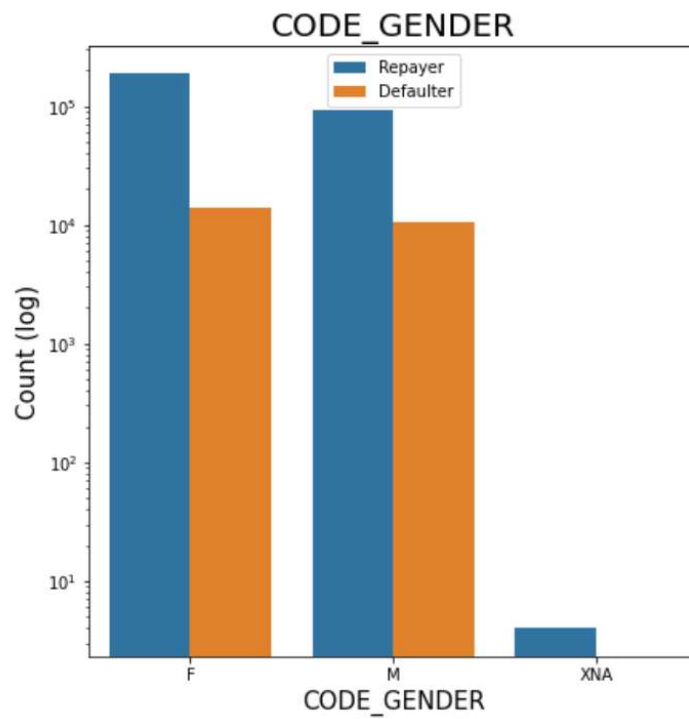
3. Own Real State

`Inferences:`

- The clients who own real estate are more than double of the ones that don't own.
- The defaulting rate of both categories are around the same (~8%). Thus we can infer that there is no correlation between owning a reality and defaulting the loan.
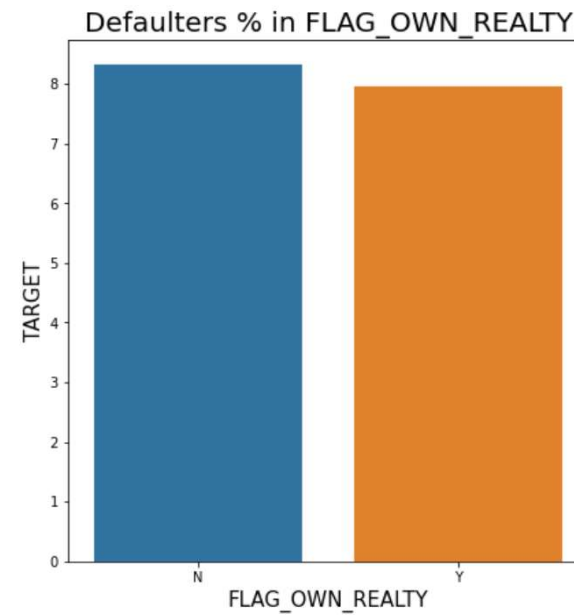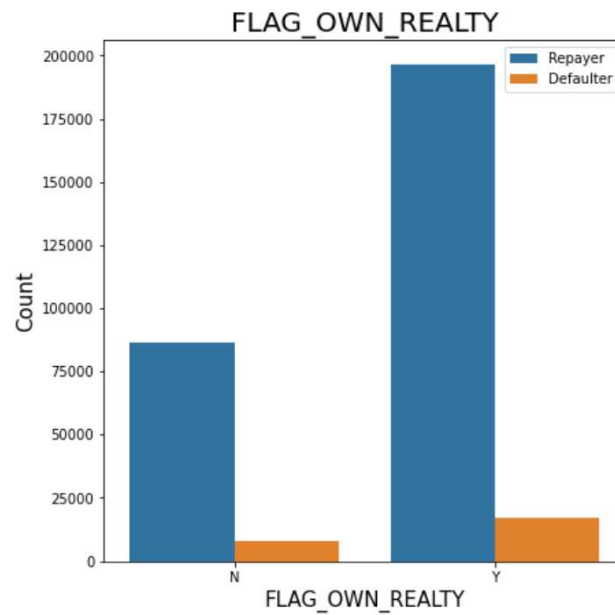
4. Housing Type

Inferences:

- Majority of people live in House/apartment.
- People living in office apartments have lowest default rate.
- People living with parents (~11.5%) and in rented apartments(>12%) have higher probability of defaulting.
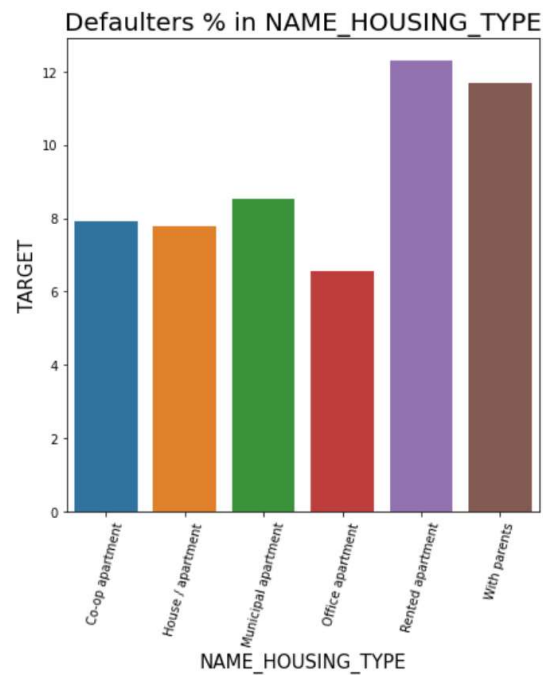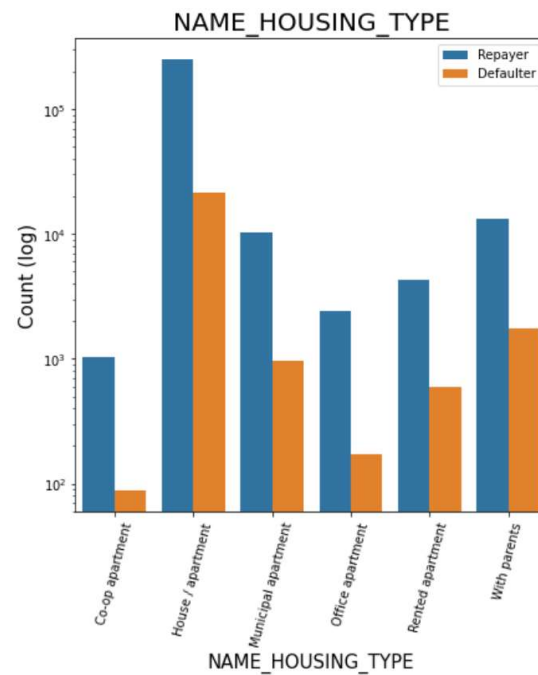
5. Family Status

Inferences:

- Most of the people who have taken loan are married, followed by Single/not married and civil marriage.
- In Percentage of defaulters, Civil marriage and Single/not married has the highest percent around (10%) and widow has the lowest around 6% (exception being Unknown).

6. Education Type

- Majority of clients have Secondary/secondary special education, followed by clients with Higher education.
- Very few clients have an academic degree
- Lower secondary category client have highest rate of defaulting around 11%.
- People with Academic degree are least likely to default.
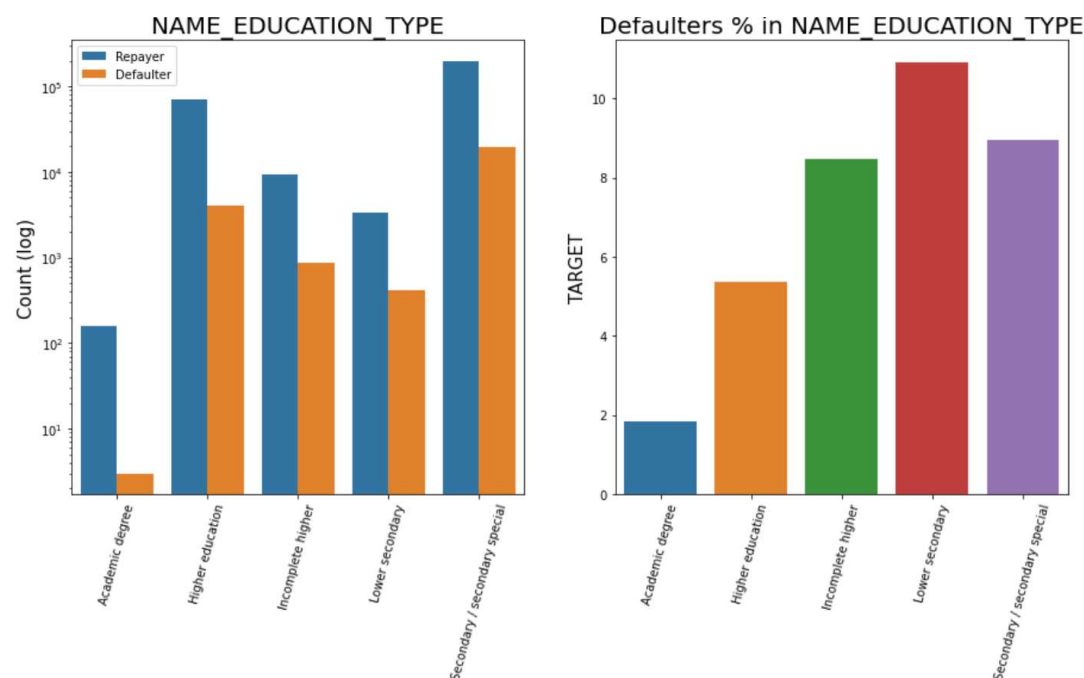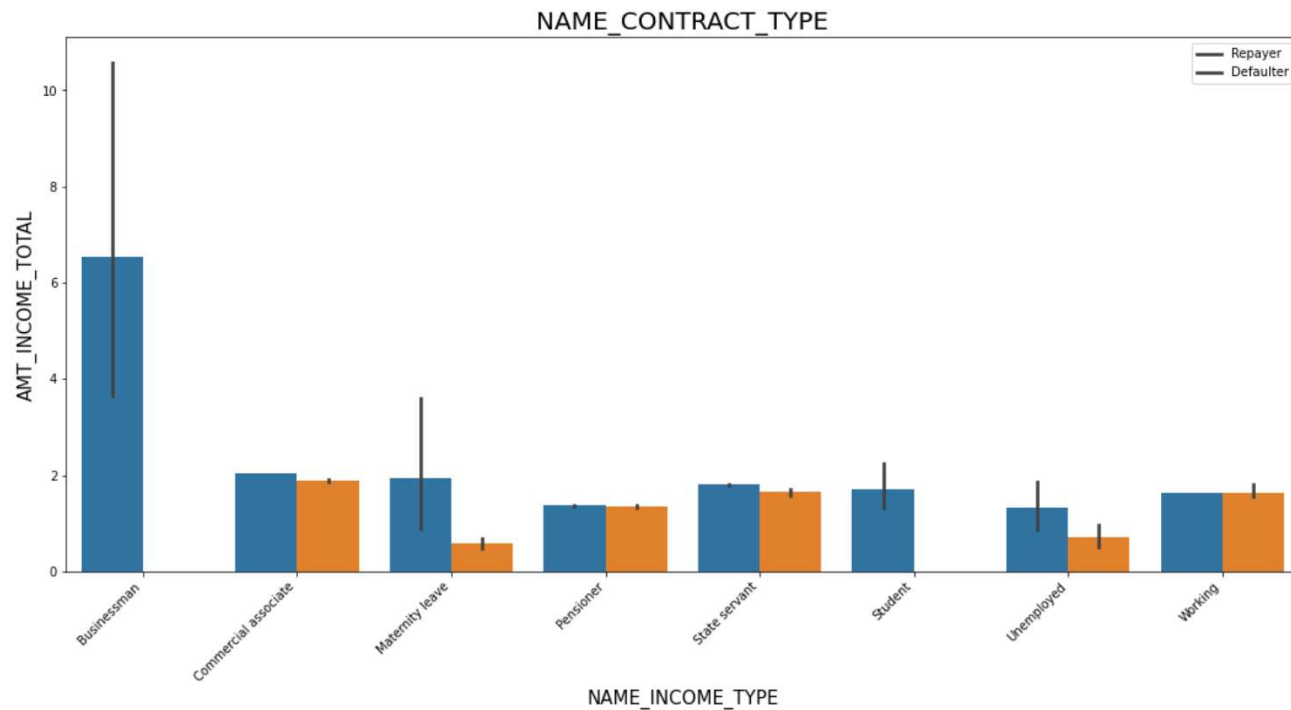
7. Income Type

- Most of applicants for loans income type is Working, followed by Commercial associate, Pensioner and State servant.
- The applicants who are on Maternity leave have defaulting percentage of 40% which is the highest, followed by Unemployed (37%). The rest categories have under 10% defaultees.
- Student and Businessmen though less in numbers, do not have default record. Safest two categories for providing loan.

UpGrad
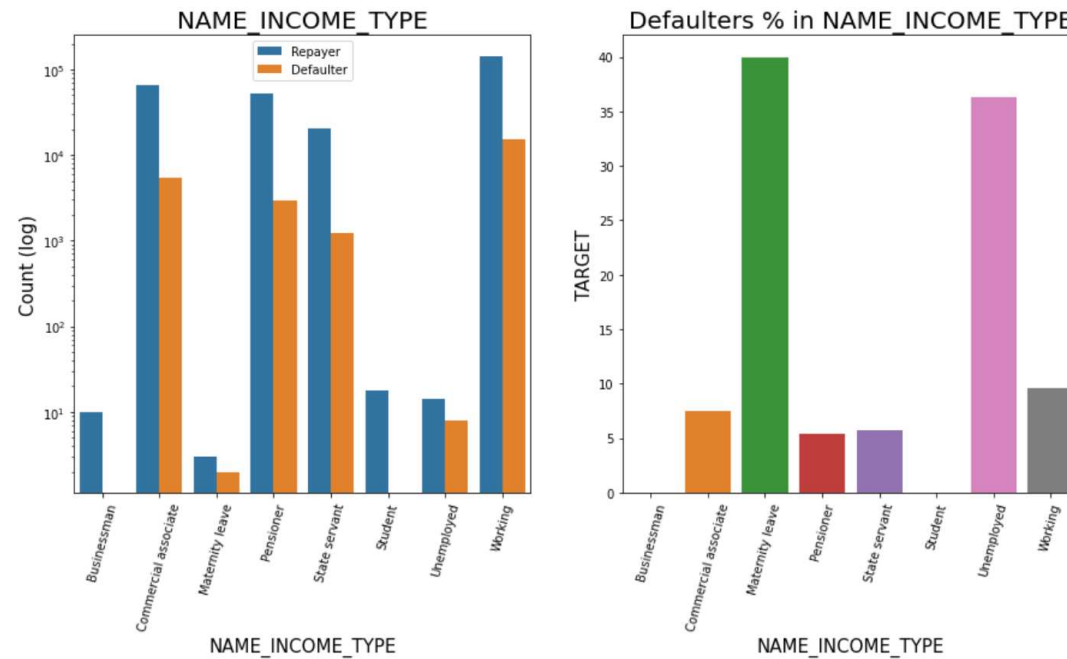
Income Type Vs Income Total

Inferences:

- It can be seen that Businessman income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a Businessman could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs

Top 10 correlation for the Re-Payers

VAR1	VAR2	Correlation
AMT_GOODS_PRICE	AMT_CREDIT	0.987250
AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
AMT_ANNUITY	AMT_CREDIT	0.771309
DAYS_EMPLOYED	DAYS_BIRTH	0.626114
AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349462
AMT_CREDIT	AMT_INCOME_TOTAL	0.342799
DAYS_REGISTRATION	DAYS_BIRTH	0.333151
DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.276663
DAYS_ID_PUBLISH	DAYS_BIRTH	0.271314

Inferences:

1. Credit amount is highly correlated with:

- Goods Price Amount
- Loan Annuity
- Total Income

2. We can also see that repayers have high correlation in number of days employed.

Top 10 correlation for the Defaulter

Inferences:

- Credit amount is highly correlated with good price amount which is same as repayers.
- Loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayers(0.77)
- We can also see that repayers have high correlation in number of days employed(0.62) when compared to defaulters(0.58).
- There is a severe drop in the correlation between total income of the client and the credit amount(0.038) amongst defaulters whereas it is 0.342 among repayers.
- Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in repayers.
- There is a slight increase in defaulted to observed count in social circle among defaulters(0.264) when compared to repayers(0.254)

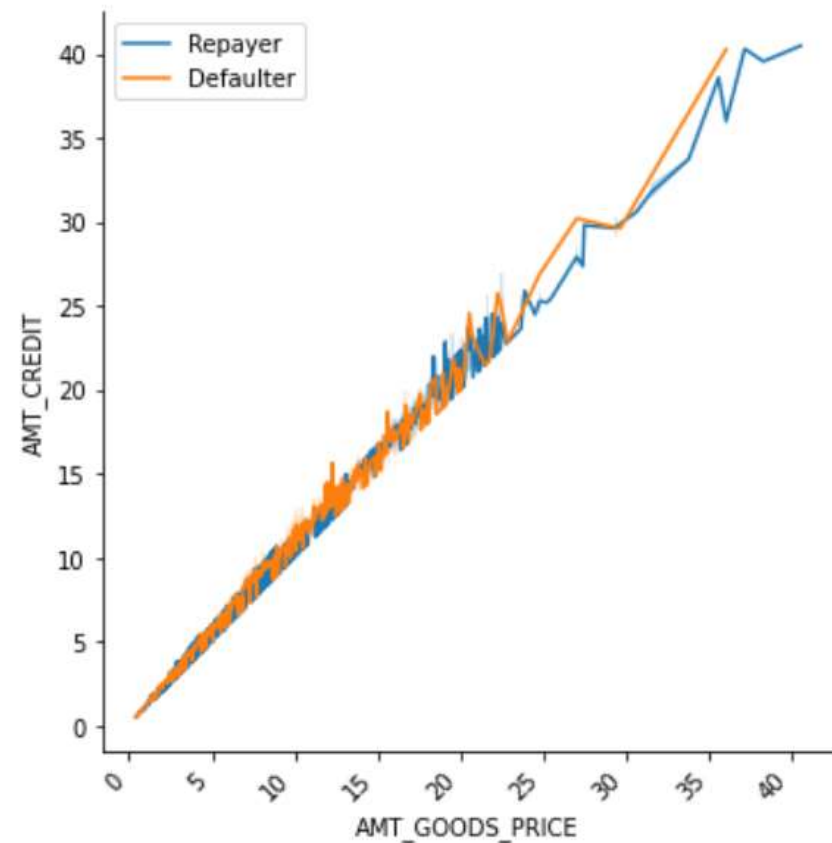| VAR1	VAR2	Correlation
AMT_GOODS_PRICE	AMT_CREDIT	0.983103
AMT_GOODS_PRICE	AMT_ANNUITY	0.752699
AMT_ANNUITY	AMT_CREDIT	0.752195
DAYS_EMPLOYED	DAYS_BIRTH	0.582185
DAYS_REGISTRATION	DAYS_BIRTH	0.289114
FLAG_DOCUMENT_3	DAYS_EMPLOYED	0.272169
DEF_60_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.264159
DAYS_ID_PUBLISH	DAYS_BIRTH	0.252863
AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_HOUR	0.247511
DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.229090

Amount Vs Density

Inferences:

- Most no of loans are given for goods price below 10 lakhs
- Most people pay annuity below 50K for the credit loan
- Credit amount of the loan is mostly less then 10 lakhs
- The repayers and defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision

UpGrad
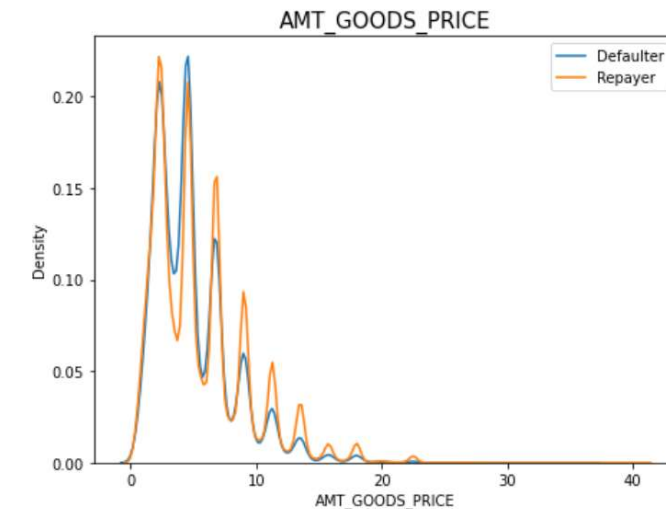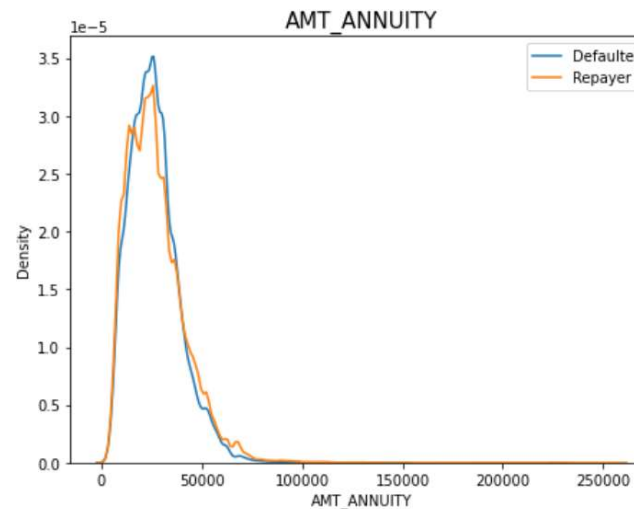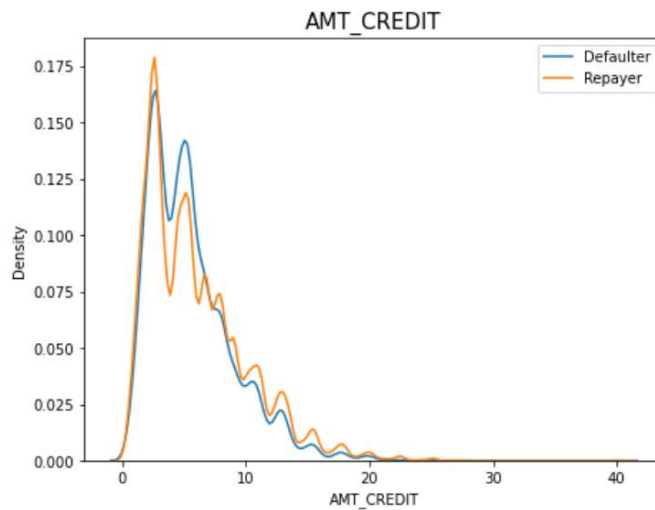
Goods price and credit and comparing with loan repayment status

Inferences:

When the credit amount goes beyond 30 Lakhs,
there is an increase in defaulters.

After analysing the datasets, there are few attributes of a client with which the bank would be able to identify if they will repay the loan or not. The analysis is consisted as below with the contributing factors and categorization:

A. Decisive Factor whether an applicant will be Repayer:

1. NAME_EDUCATION_TYPE: Academic degree has less defaults.

2. NAME_INCOME_TYPE: Student and Businessmen have no defaults.

3. REGION_RATING_CLIENT: RATING 1 is safer.

4. ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%

5. DAYS_BIRTH: People above age of 50 have low probability of defaulting

6. DAYS_EMPLOYED: Clients with 40+ year experience having less than 1% default rate

7. AMT_INCOME_TOTAL: Applicant with Income more than 700,000 are less likely to default

8. NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, Buying garage are being repayed mostly.

9. CNT_CHILDREN: People with zero to two children tend to repay the loans.

B. Decisive Factor whether an applicant will be Defaulter:

1. CODE_GENDER: Men are at relatively higher default rate

2. NAME_FAMILY_STATUS : People who have civil marriage or who are single default a lot.

3. NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education

4. NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot.

5. REGION_RATING_CLIENT: People who live in Rating 3 has highest defaults.

6. OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as their default rate is huge.

7. ORGANIZATION_TYPE: Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed people have relative high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.

8. DAYS_BIRTH: Avoid young people who are in age group of 20-40 as they have higher probability of defaulting

9. DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.

10. CNT_CHILDREN & CNT_FAM_MEMBERS: Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected.

11. AMT_GOODS_PRICE: When the credit amount goes beyond 3lakhs, there is an increase in defaulters.

C. Factors that Loan can be given on Condition of High Interest rate to mitigate any default risk leading to business loss:

1. NAME_HOUSING_TYPE: High number of loan applications are from the category of people who live in Rented apartments & living with parents and hence offering the loan would mitigate the loss if any of those default.

2. AMT_CREDIT: People who get loan for 3-6 Lakhs tend to default more than others and hence having higher interest specifically for this credit range would be ideal.

3. AMT_INCOME: Since 90% of the applications have Income total less than 3Lakhs and they have high probability of defaulting, they could be offered loan with higher interest compared to other income category.

4. CNT_CHILDREN & CNT_FAM_MEMBERS: Clients who have 4 to 8 children has a very high default rate and hence higher interest should be imposed on their loans.

5. NAME_CASH_LOAN_PURPOSE: Loan taken for the purpose of Repairs seems to have highest default rate. A very high number applications have been rejected by bank or refused by client in previous applications as well which has purpose as repair or other. This shows that purpose repair is taken as high risk by bank and either they are rejected, or bank offers very high loan interest rate which is not feasible by the clients, thus they refuse the loan. The same approach could be followed in future as well.

D. Suggestions:

1. 90% of the previously cancelled client have actually repayed the loan. Record the reason for cancellation which might help the bank to determine and negotiate terms with these repaying customers in future for increase business opportunity.

2. 88% of the clients who were refused by bank for loan earlier have now turned into a repaying client. Hence documenting the reason for rejection could mitigate the business loss and these clients could be contacted for further loans.