Module 8: Credit EDA Assignment

Credit EDA Case Study

An Entry into Data Analysis

Submitted by: **SHUBHAM B. KANHEKAR**
Batch : DSC73 (Oct’24)
Date: 31/12/2024

Goal of the Analysis

The goal of the Credit EDA Assignment is to get us acquainted with various cleaning and analysis techniques while also gathering Insights from the dataset provided to us. In this presentation, we will be looking at some of the many plots created during the analysis and a brief note about each plot has be added with the figures. Also, the recommended variables which should be observed by the financial institute/bank are listed in the second to last slide.

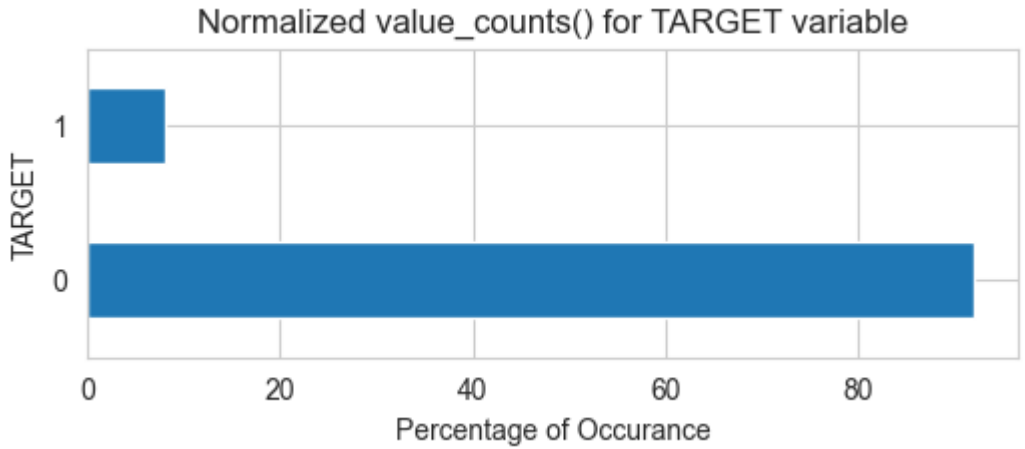


fig_1

fig_1: 'TARGET' variable value counts show that 8.07% of applicants have trouble with payments

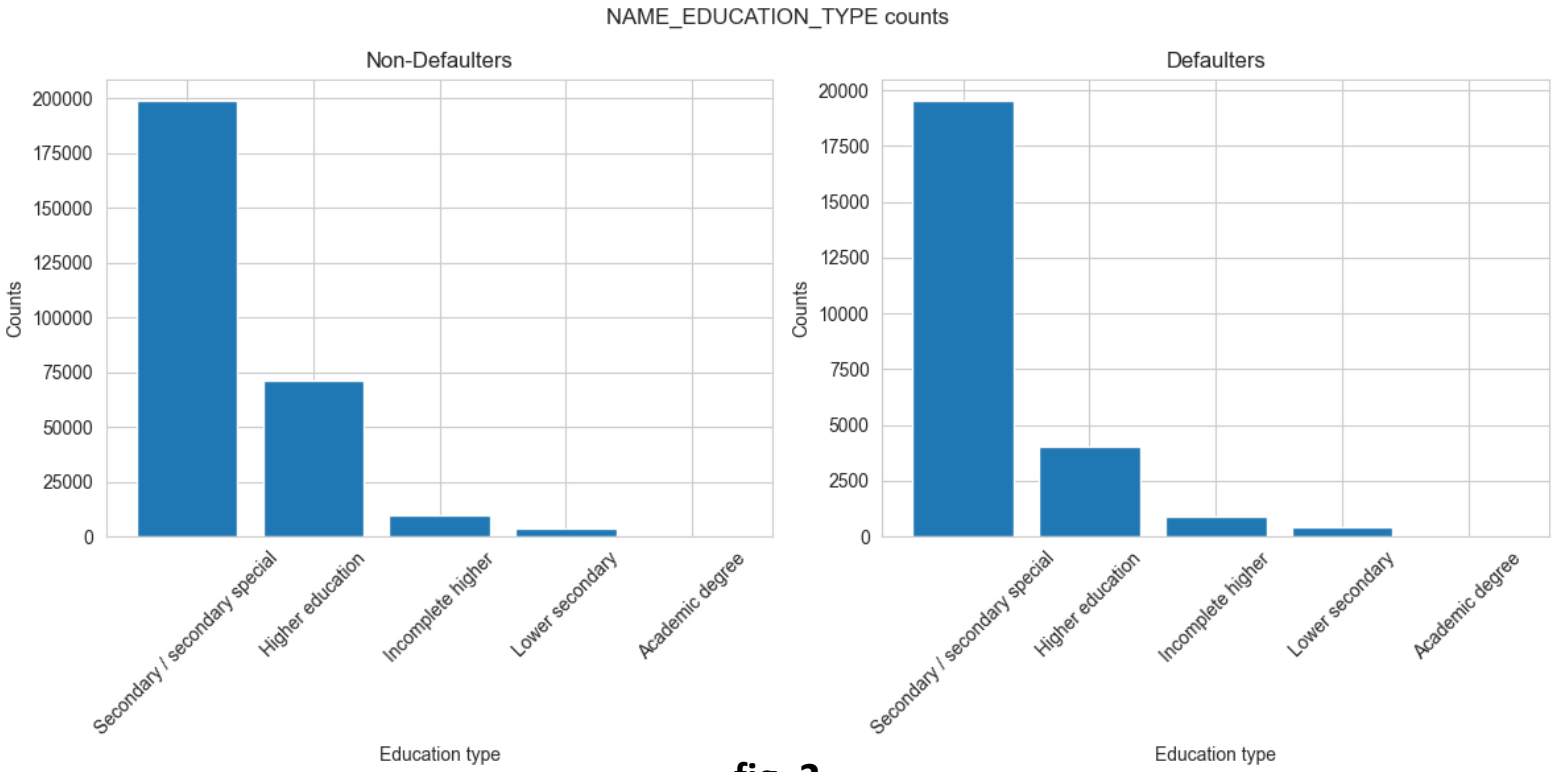


fig_2

fig_2:'Education type: Higher education' shows the most promising clients who are least probable to default on payments as we can see in fig_2; All bars are same height (proportion) except 'Higher Education' has lower proportion in defaulters plot

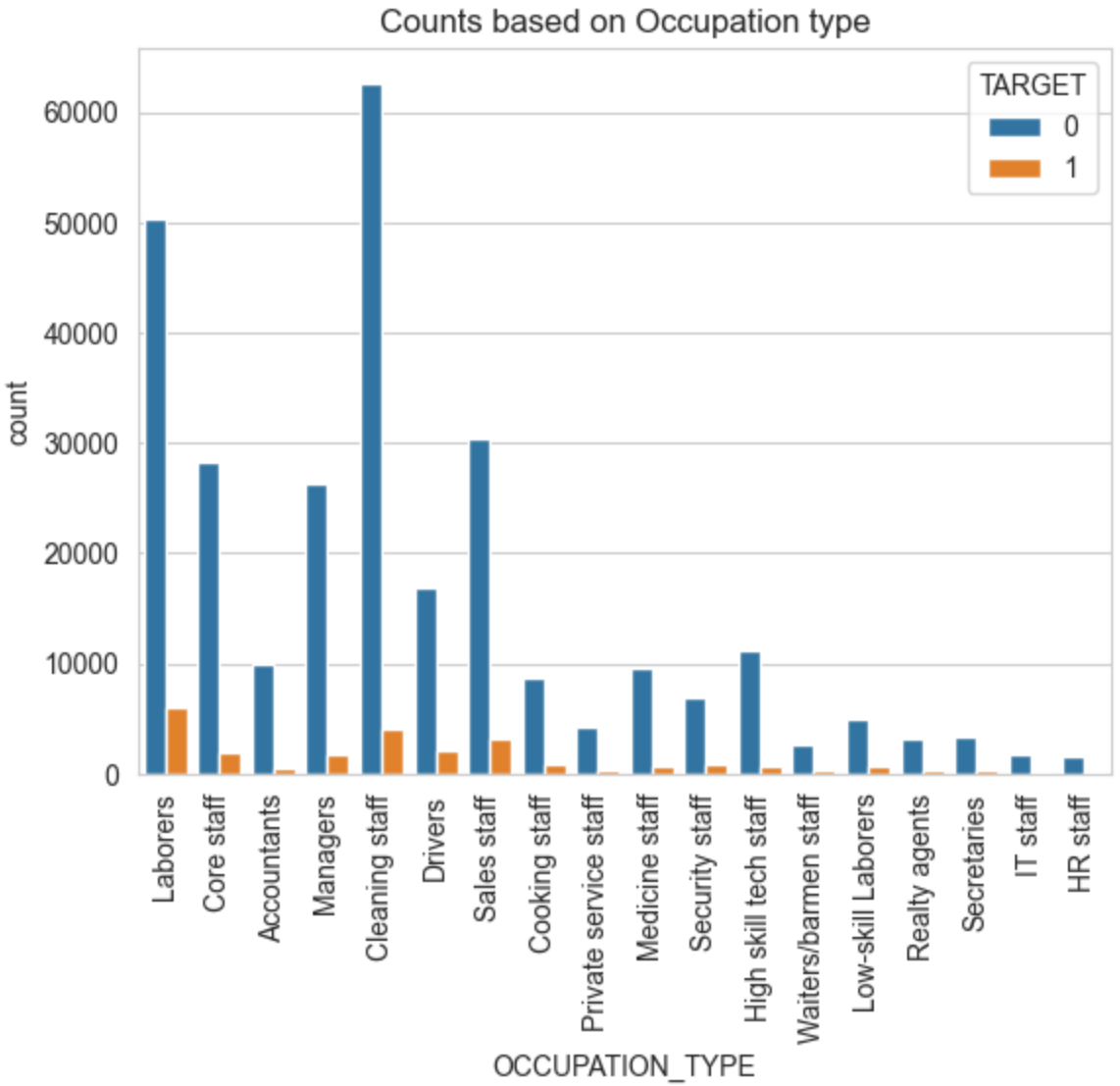


Fig_3

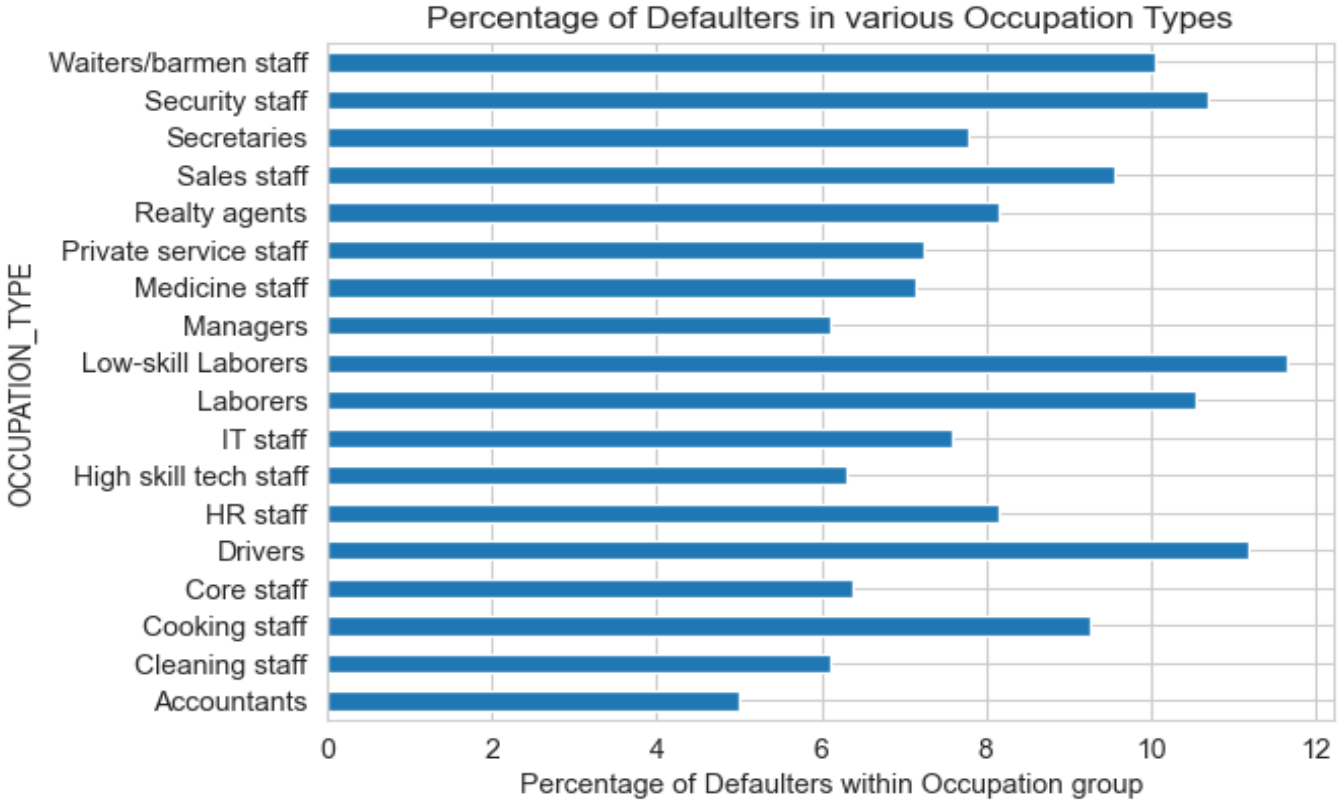


Fig_4

Fig_3 and fig_4 are essentially showing that cleaning staff have most number of occurrences but Laborers, Low-skilled Laborers, Drivers and Security staff are categories which have most defaulters percentage within their Occupation category.

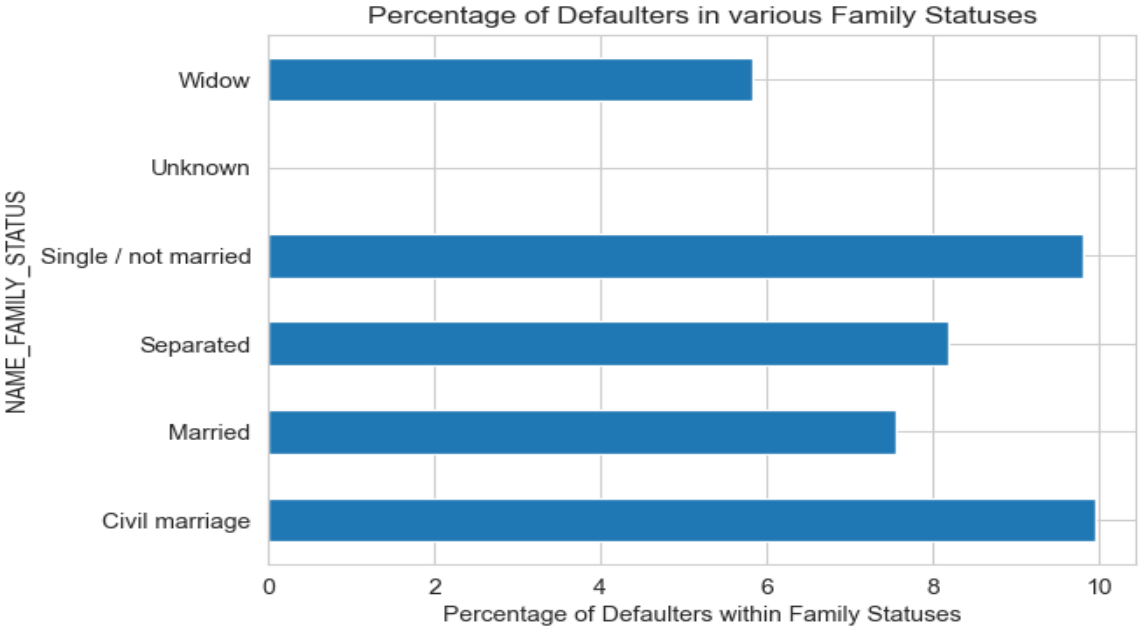


Fig_5

Fig_5 shows that single and civil marriage categories of Family statuses exhibit most probability to default on payments

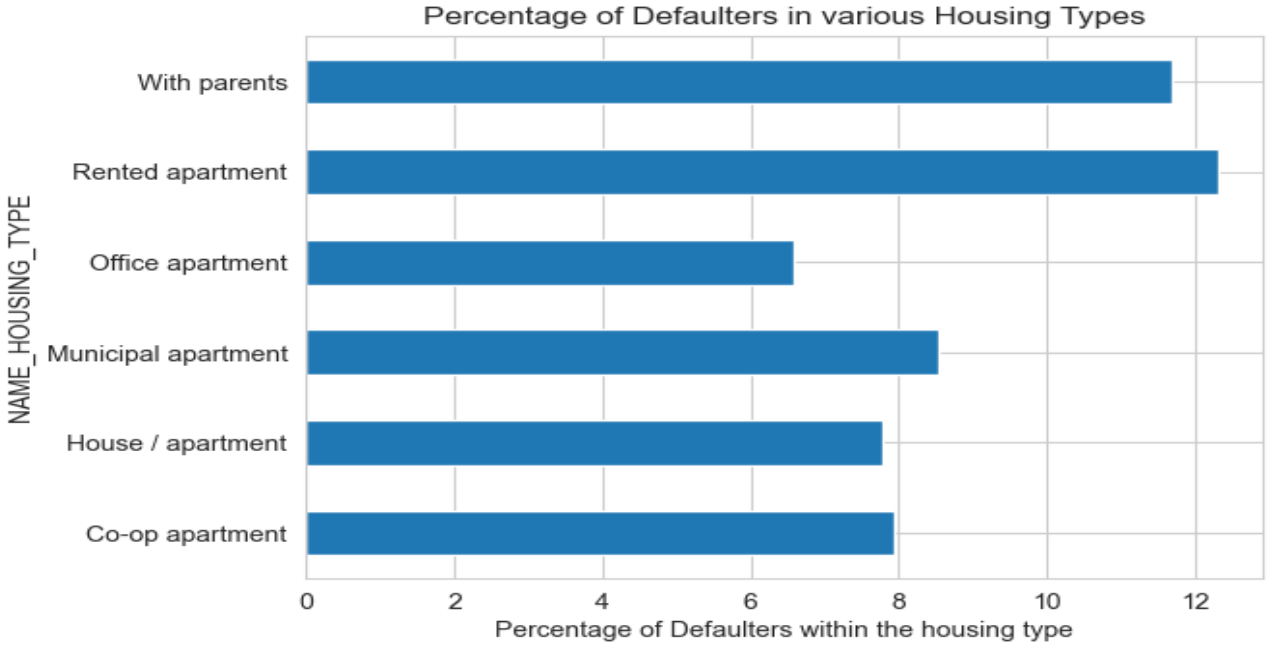


Fig_6

Fig_6 shows that People living with parents and rented apartments are most likely to default, while those living in Office apartment are least probable to default

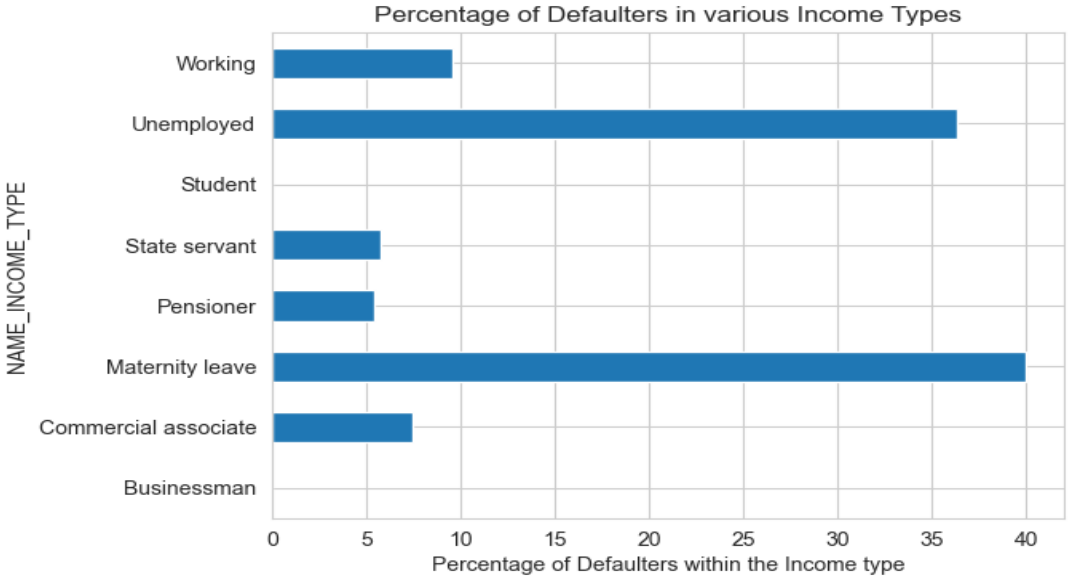
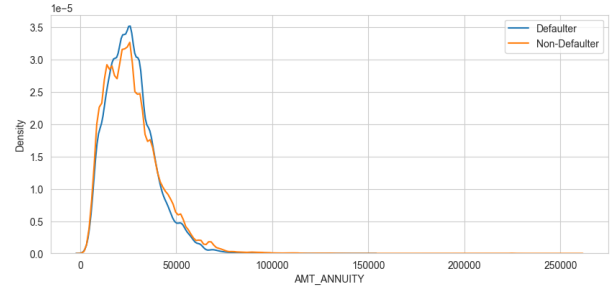
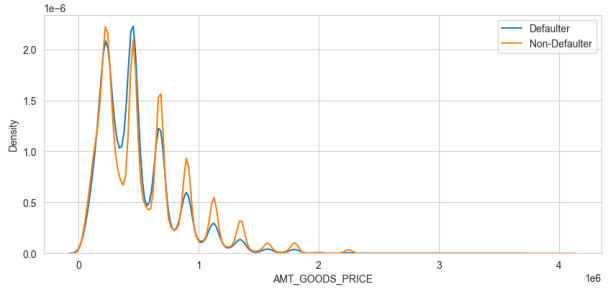
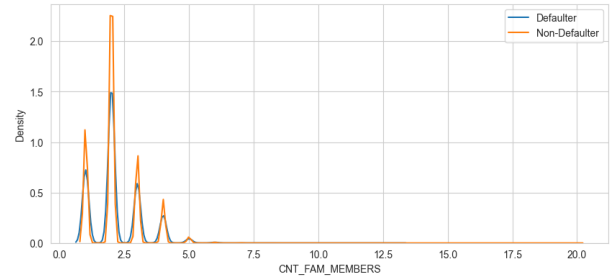


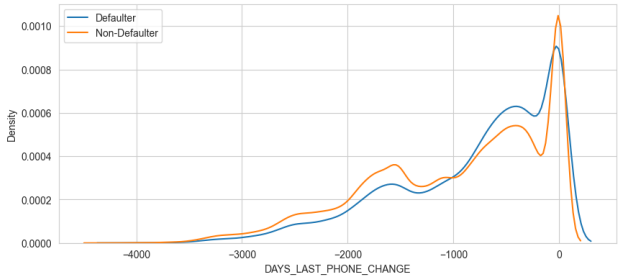
Fig_7

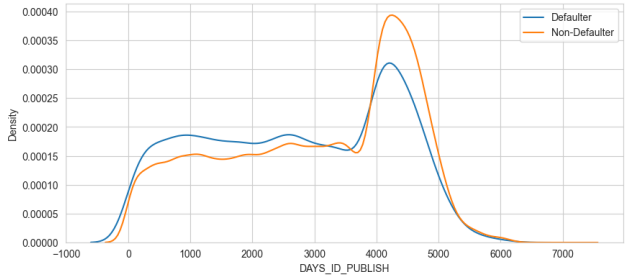
fig_7 : Unemployed and maternity leave categories are most probable to have payment difficulties while Pensioners, State servant,Commercial associate and working Income Types have least probability to default

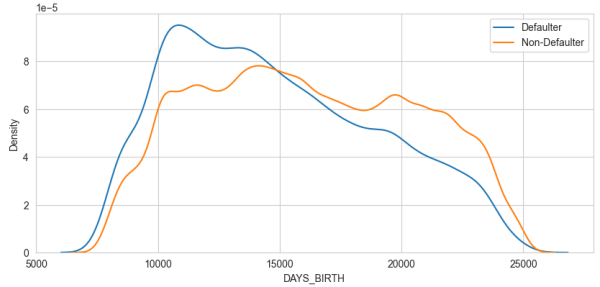


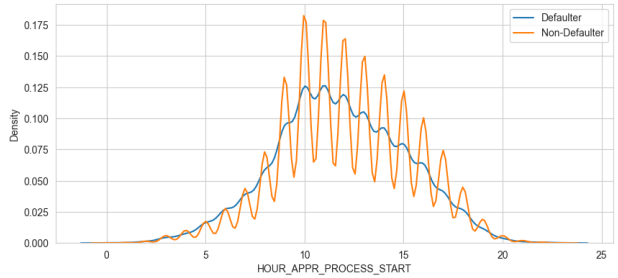


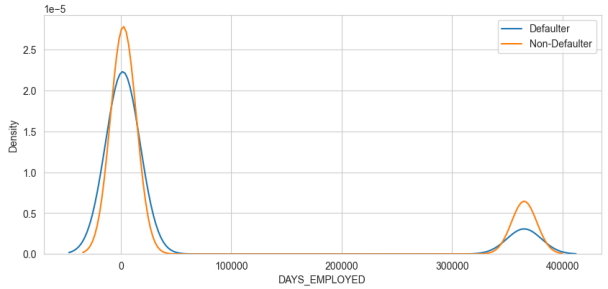


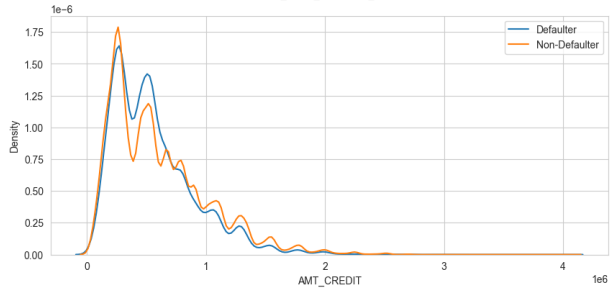












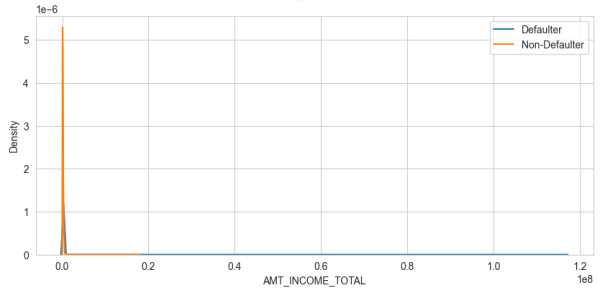


Fig_8: 10 continous numerical columns:
'AMT_ANNUITY','AMT_GOODS_PRICE','CNT_FAM_MEMBERS', 'DAYS_LAST_PHONE_CHANGE', 'DAYS_ID_PUBLISH', 'DAYS_BIRTH', 'HOUR_APPR_PROCESS_START', 'DAYS_EMPLOYED', 'AMT_CREDIT', 'AMT_INCOME_TOTAL'

Insights:

- A significant number of applications are submitted between 9 AM and 2 PM in both the Current and Previous datasets, indicating that these are the bank's busiest hours.
- Families of size 1 to 4 (i.e. nuclear families) tend to apply for more loans compared to other family types.
- Most of the applicants received lower credits based on distribution of 'AMT_CREDIT' variable
- Looking at 'DAYS_EMPLOYED's kdeplot, applicants are either at the starting of their careers or are very senior professionals.
- from 'DAYS_BIRTH' we can say that younger people are more likely to default while older applicants defaulted lesser.
- 'AMT_ANNUITY' kdeplot shows that lower annuity shows more possibility of defaulting.

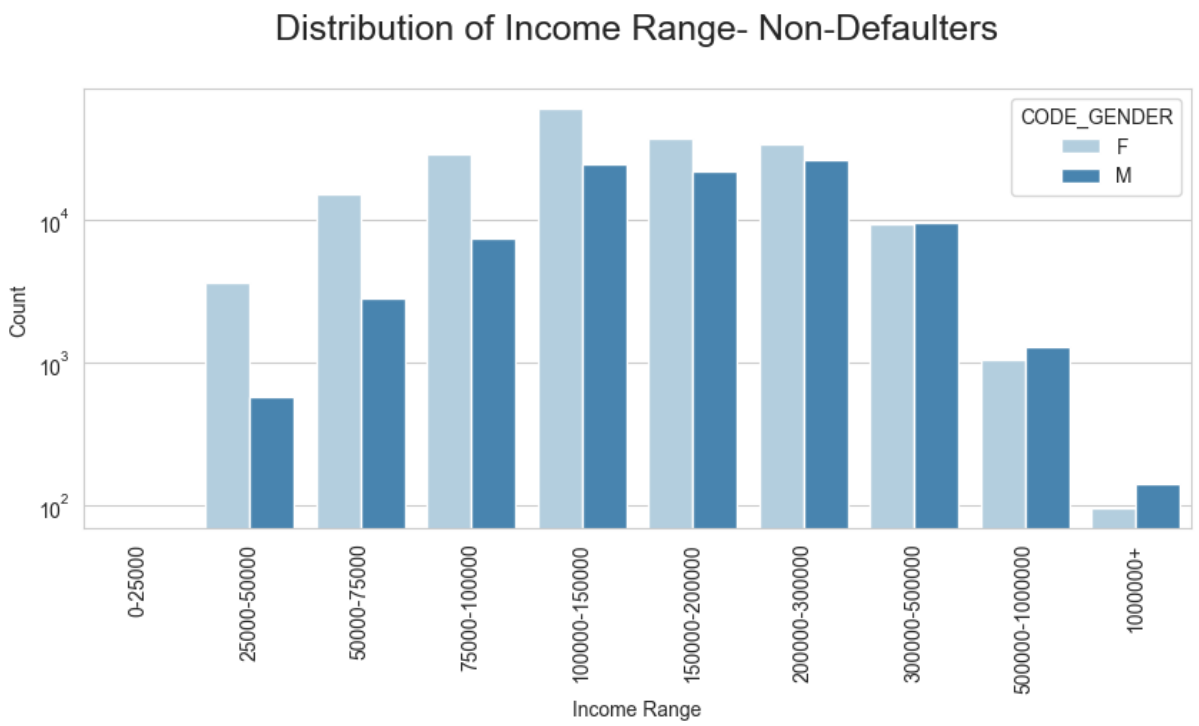


Fig_9

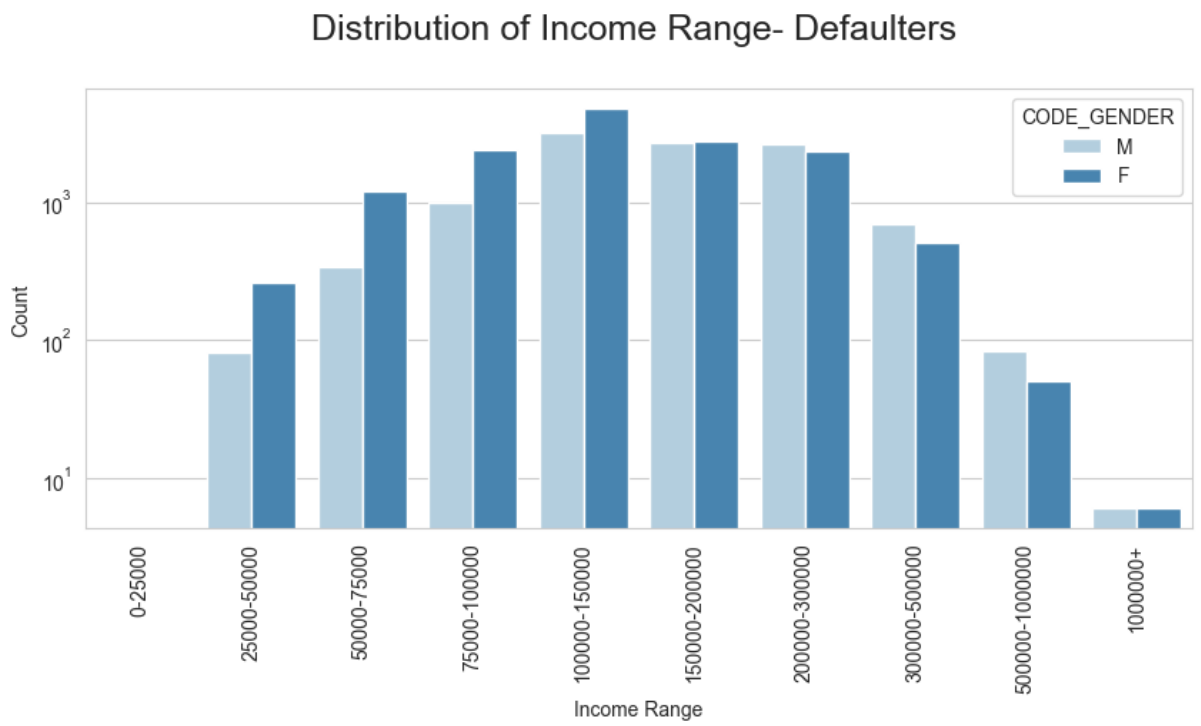
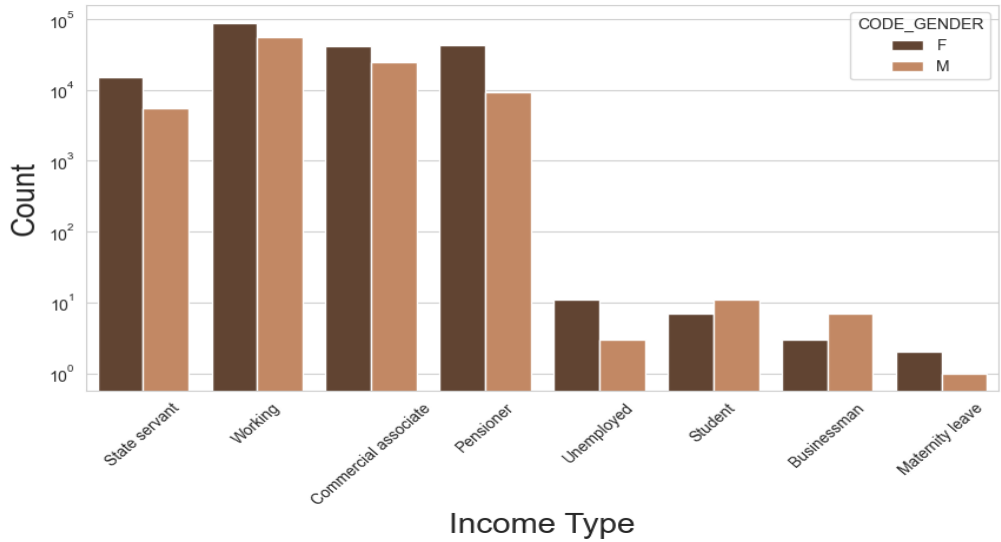


Fig_10

Fig_9 and fig_10 show that most applicants are from income range of 50000 to 500000 and most common category is 100000-150000 in case of both defaulters and non- defaulters. Also there are more Females than males in almost all categories

Fig_11 Distribution of Income Type



In Fig_11, State servant, working, Commercial associate and Pensioner are the most common Income types among applicants. And Females dominate the number of applications in all these four Income types

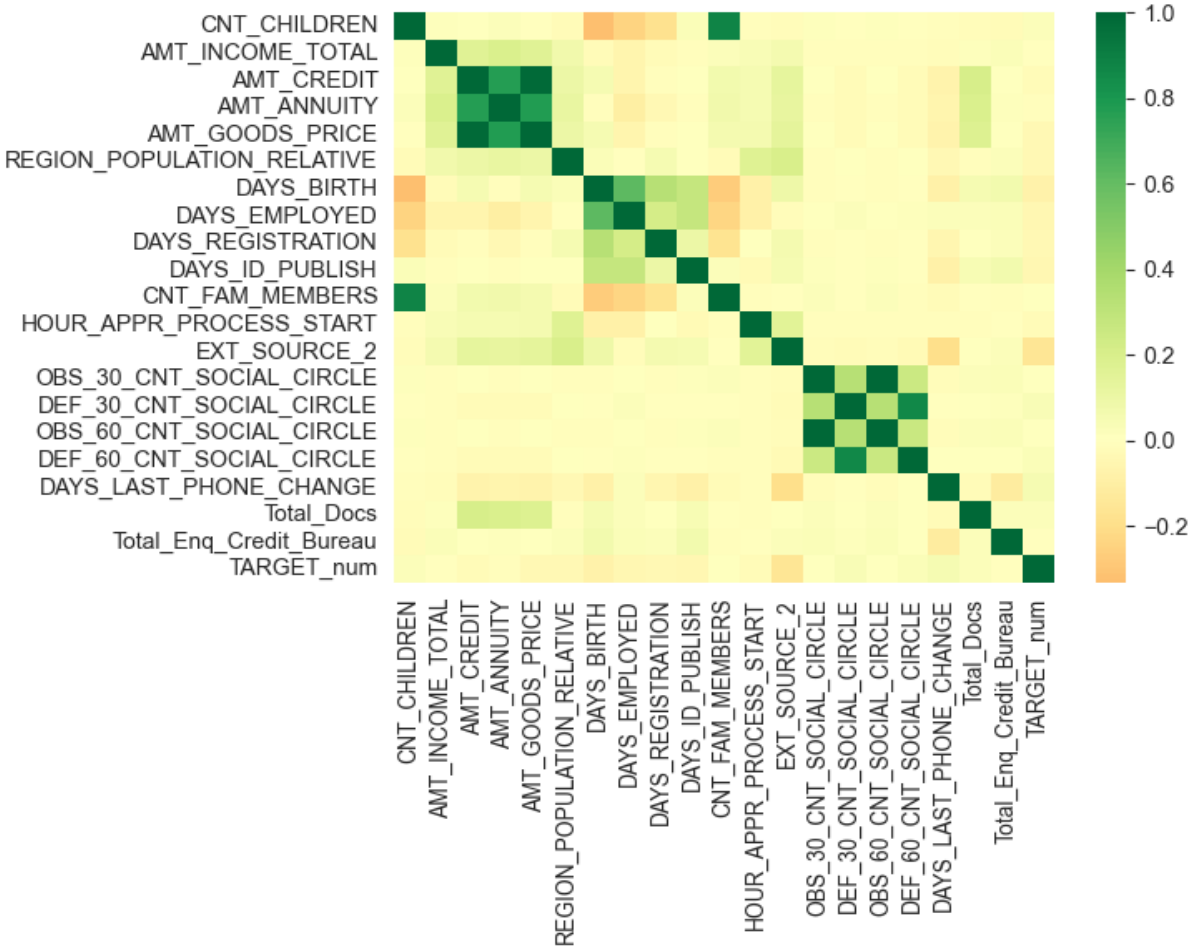
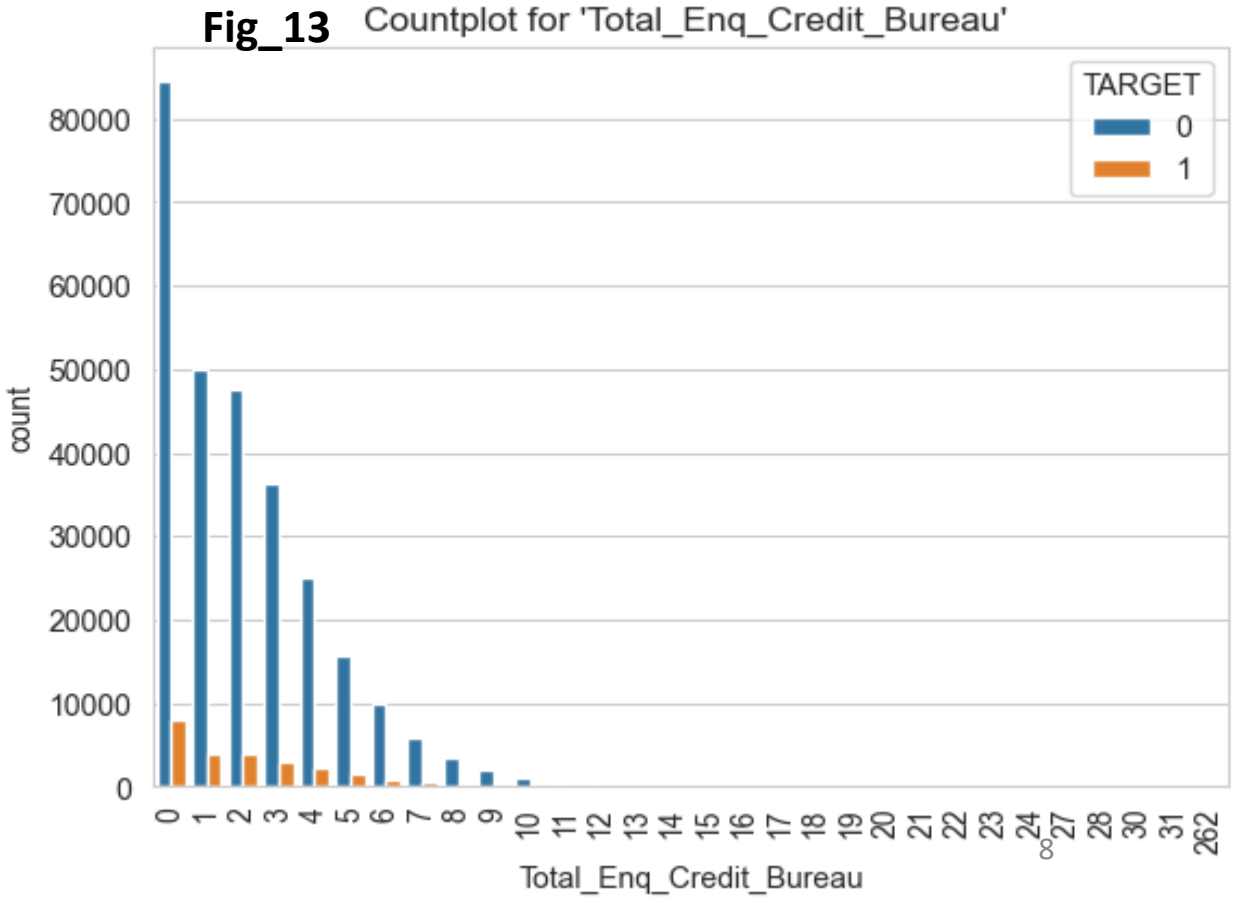


Fig_12

In fig_12, we can observe that correlation is there between the following:
strong correlation: CNT_FAM_MEMBERS vs CNT_CHILDREN, OBS_60_CNT_SOCIAL_CIRCLE vs OBS_30_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE vs DEF_30_CNT_SOCIAL_CIRCLE, DAYS_BIRTH vs DAYS_EMPLOYED, AMT_ANNUITY vs AMT_CREDIT, AMT_GOODS_PRICE vs AMT_CREDIT, AMT_GOODS_PRICE vs AMT_ANNUITY
weak correlation: DAYS_BIRTH vs CNT_CHILDREN, TARGET vs EXT_SOURCE_2

Fig_13



In fig_13, we observe that most of the applicants have lesser number of credit bureau enquiries

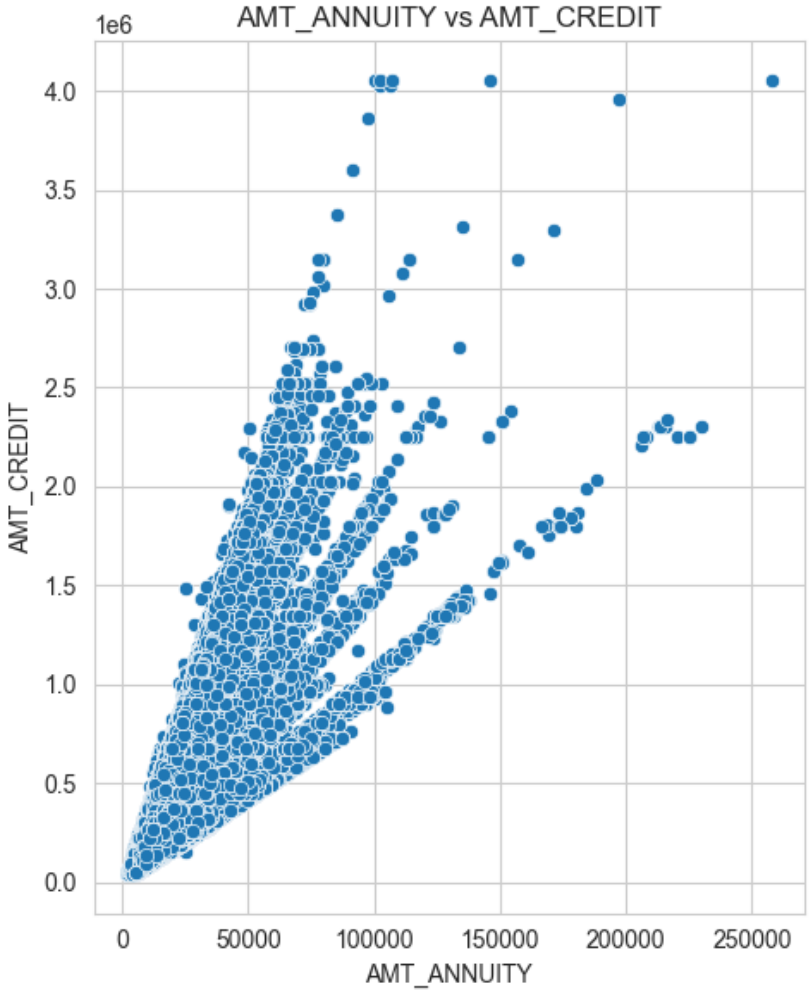


Fig_14

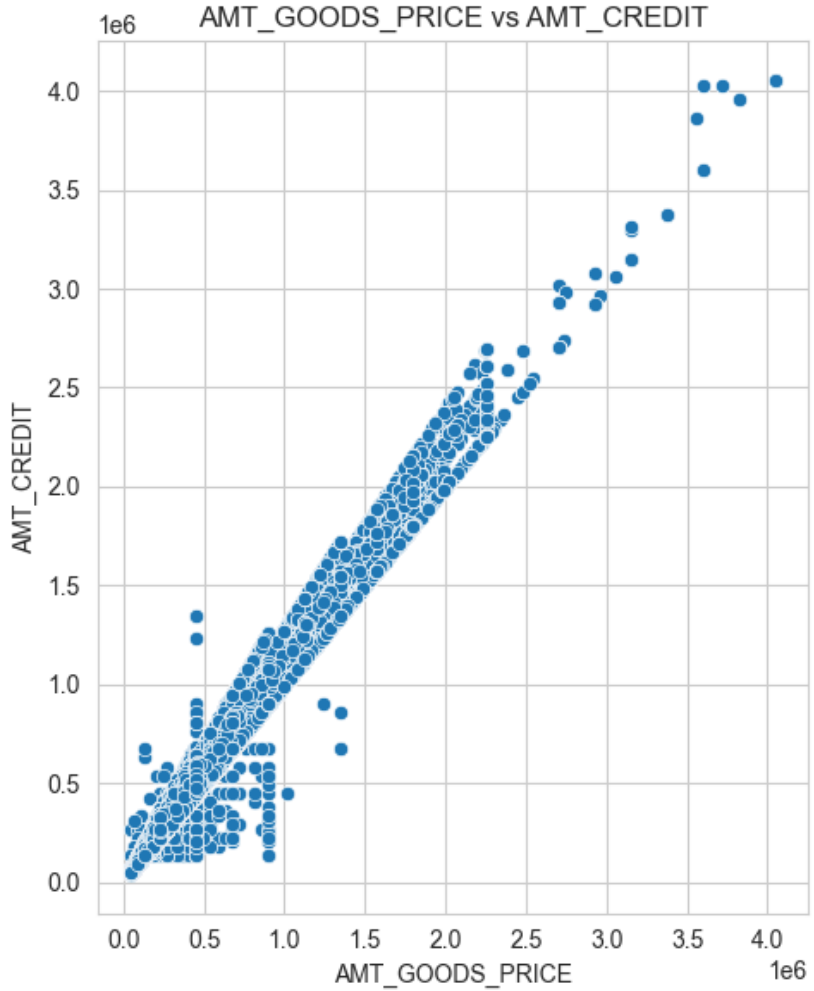


Fig_15

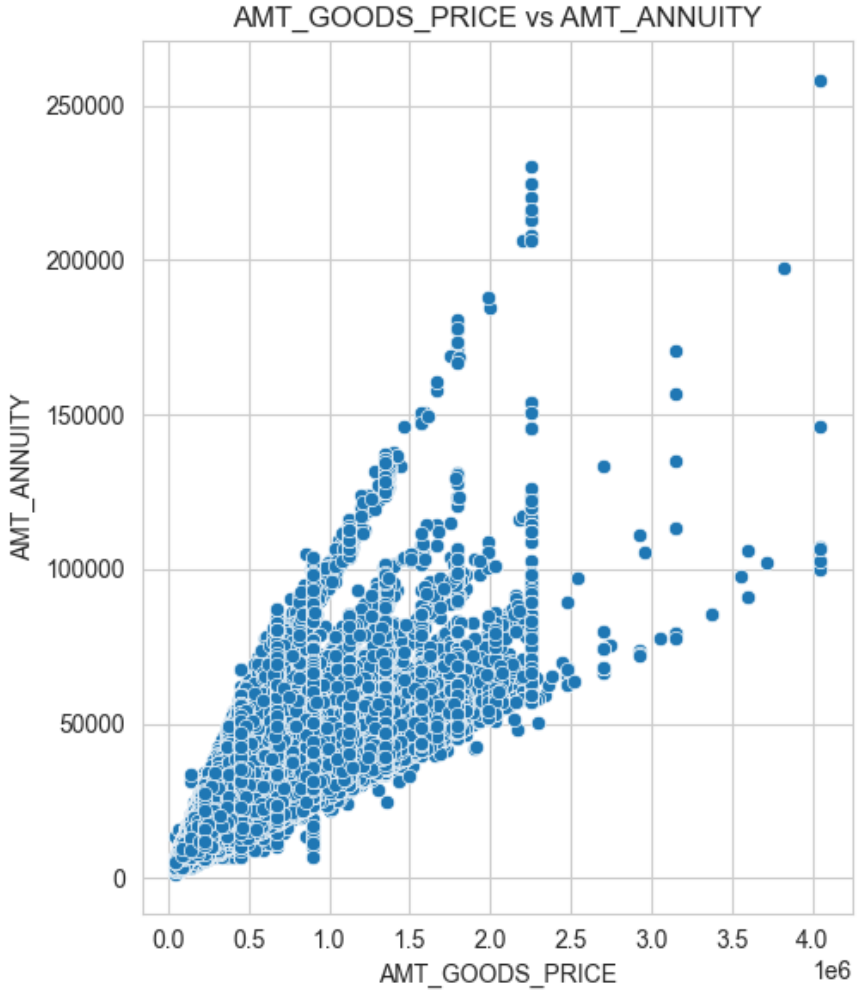


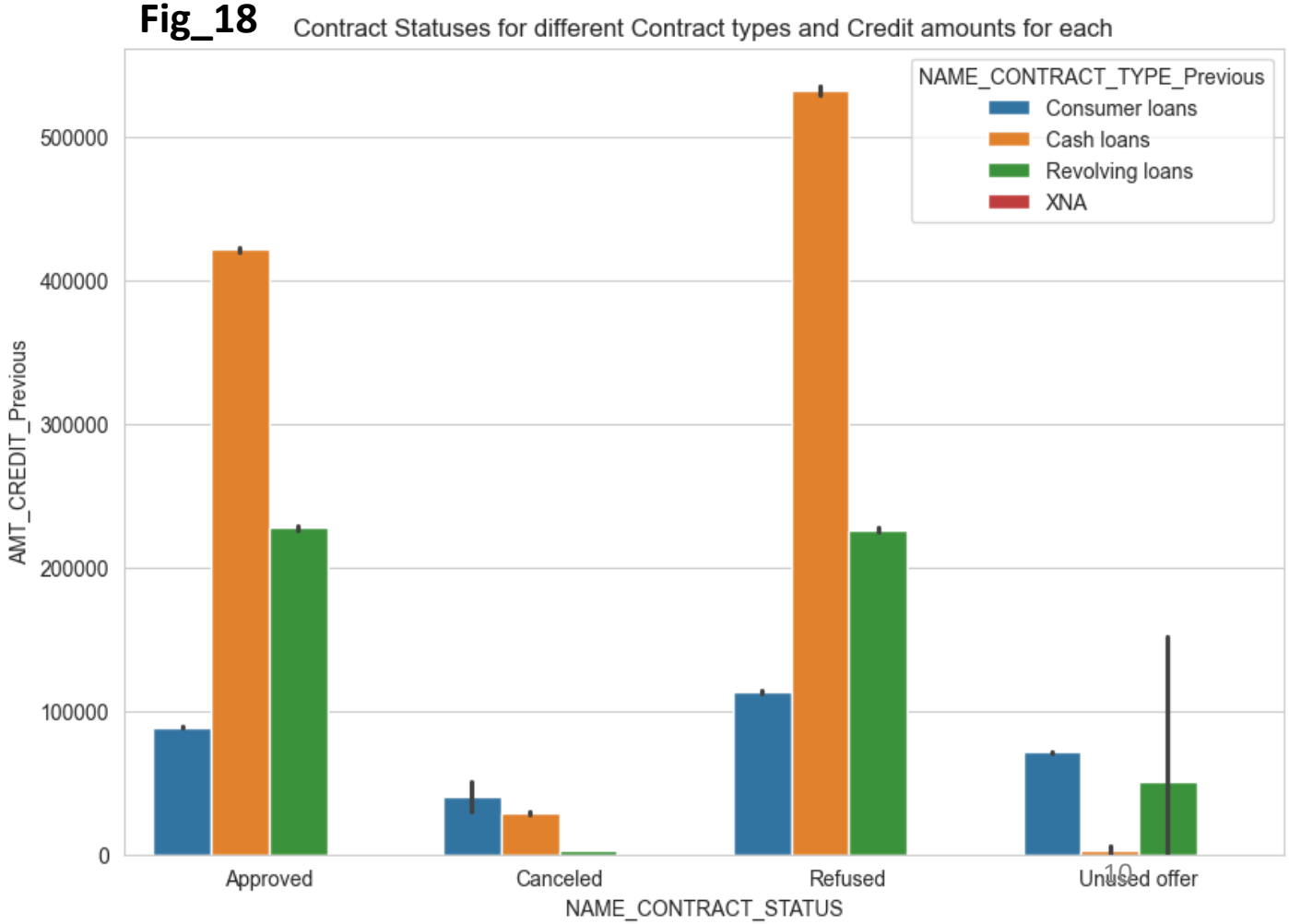
Fig_16

Fig_14, Fig_15 and Fig_16 show that the variables: 'AMT_ANNUITY', 'AMT_CREDIT', 'AMT_GOODS_PRICE' are hugely correlated as one of them increases other two also show similar increase in values.

Fig_17

In Fig_17, we clearly see that a higher total income increases the probability of having higher credit amount

Fig_18



In fig_18, we observe that more cash loans are more refused than approved. Also Revolving loans have amost 50% chance of getting approved or refused. Revolving loan applications are rarely cancelled. Cash loans are engaged with most amount of credit

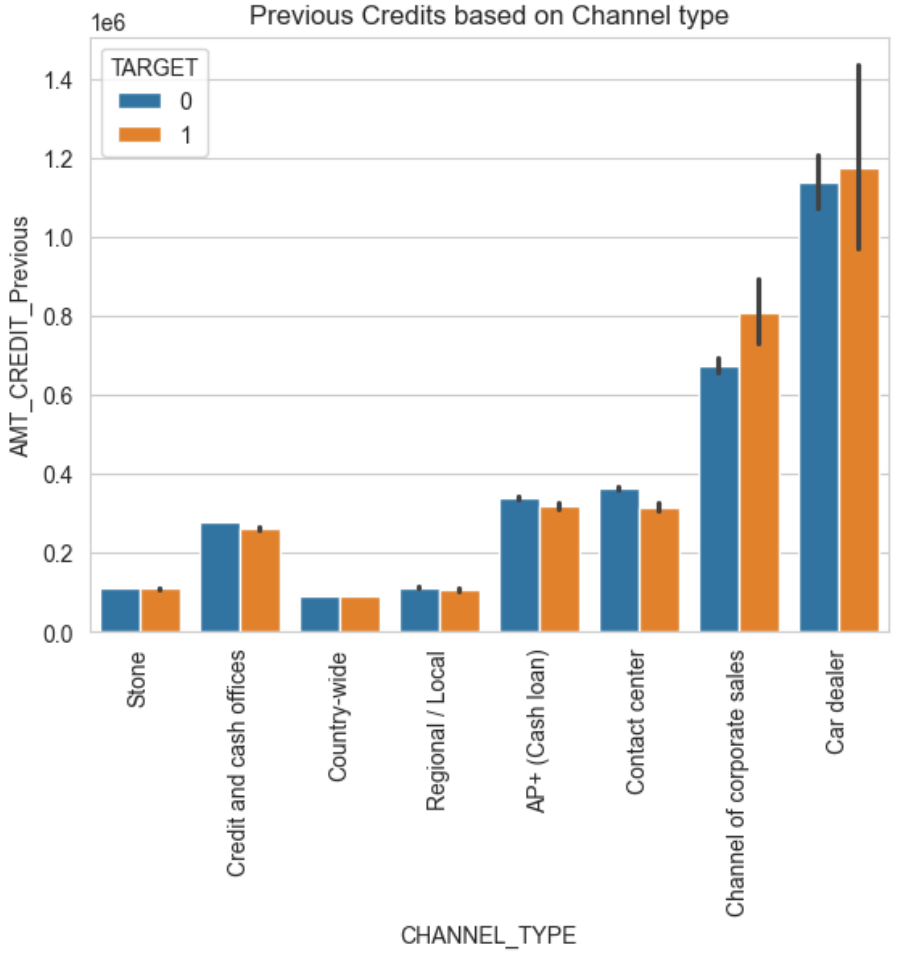


Fig_19

In fig_19,we observe that car dealers and Corporate Sales Channel have the highest amount of credit, while local channel, stone and credit and cash offices had lower credit amounts in historical data

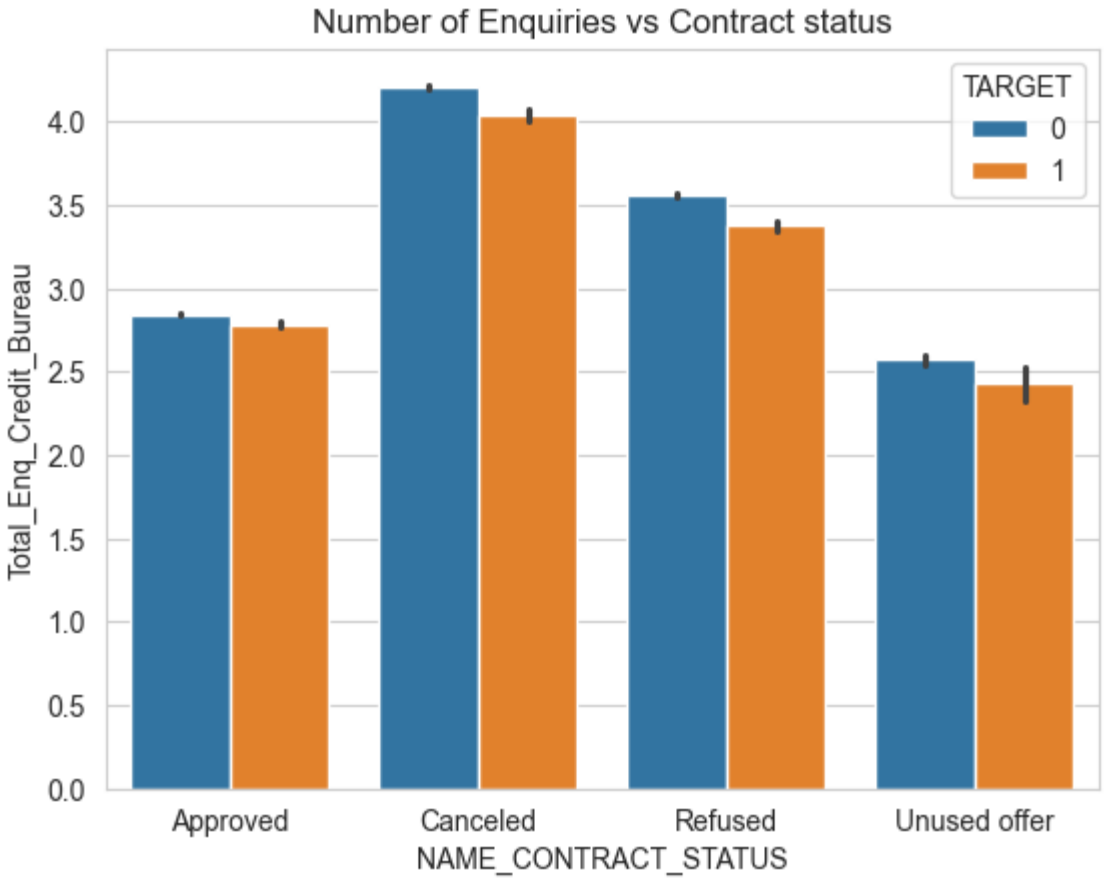


Fig_20

In fig_20, we see that more number of Credit Bureau enquiries correspond to either Cancelling or Refusing of the contracts. While Lesser number of Enquiries increases likelihood of Approval or getting (Unused) offer for the loans. However, we cannot tell much about possibility of Defaulting based on these two variables.

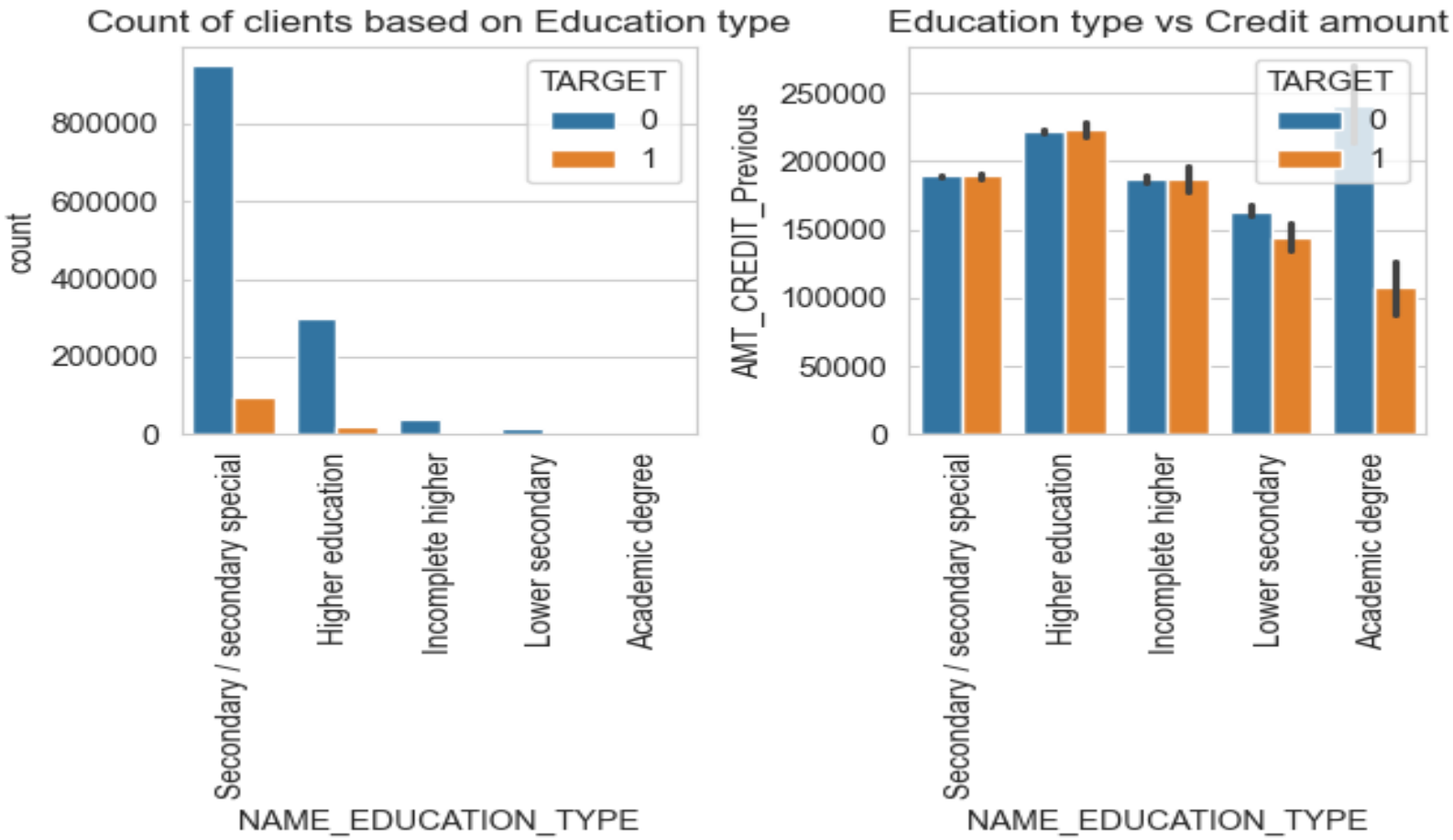


Fig_21

In Fig_21, we observe that people with higher education and secondary special education apply for most number of loans. We also see that Academic degree clients are very few but they default when credit amount is lesser. People with Higher Education are most likely to get a higher credit amount

Final Recommendations after Analysis

The following variables exhibit significant potential in predicting the likelihood of a client defaulting on payments (preferrence for non-defaulting clients):

- **AMT_INCOME_TOTAL** (higher=>non-defaulter)
- **NAME_EDUCATION_TYPE** (prefer: higher education)
- **AMT_ANNUITY** (higher => non-defaulter)
- **NAME_INCOME_TYPE** (prefer: working, commercial associate, pensioner, state servant, avoid: Unemployed, Maternity leave)
- **DAYS_EMPLOYED** (the more the better)
- **DAYS_BIRTH** (the more the better)
- **Total_Enq_Credit_Bureau** (custom column) (lesser the better)
- **CHANNEL_TYPE** (contact center has lease defaulter to non defaulters ratio)
- **OCCUPATION_TYPE** (prefer: private service, medicine, high skill teck staff, realty agents, secretaries; avoid: low-skilled laborers, laborers, cleaning staff, sales staff, drivers)
- **NAME_HOUSING_TYPE** (prefer: office apartment,house/apartment, co-op apt avoid: with parents and rented apartment)

This concludes our analysis. However, data analysis is inherently an ongoing process, and there are always additional insights that can be discovered. Therefore, the journey of exploration and understanding never truly ends.

The End