CREDIT EDA ASSIGNMENT

By KRITI ARORA
DSC 40 Batch

OBJECTIVE

This assignment aims to find trends and patterns in a bank credit data which can help the bank decide their potential customers. The target is to check what type of customers fail to repay the credit amount so that the bank can prevent issuing loan to such people.

1 BUSINESS UNDERSTANDING

The financial institutions find it hard to provide loans to people due to the missing credit history. The defaulters makes use of this advantage.
When a company receives a loan application , the company has the right to approve or reject the loan.
But there are two risks associated with this:

1) If the client is likely to pay the loan then not approving the loan tends in a business loss of the company.
2) If the client is not likely to pay the loan, then approving the loan can lead to financial loss of the company.

When a client applies for a loan, there are four types of decisions that could be taken by the bank/company:

1) Approved
2) Cancelled
3) Refused
4) Unused offer: The loan has been cancelled by the applicant but at different stages of the process.

2

STEPS TAKEN

- Importing the libraries
- Loading of datasets
- Data Inspection and Cleaning
- Handling outliers
- Analysis
  a) Univariate
  b) Bivariate
  c) Multivariate

3 IMPORTING LIBRARIES

Libraries which are used for this analysis are:

Numpy: for mathematical operations

Pandas: for handling dataframes

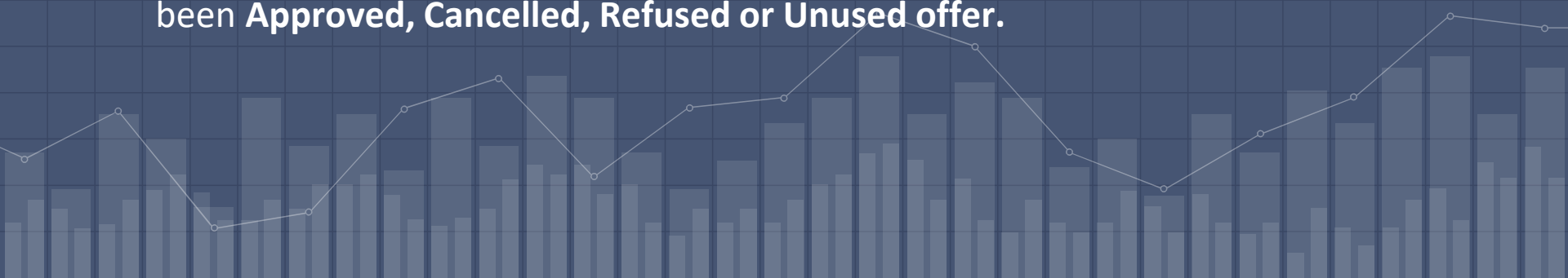
Matplotlib, Seaborn: for plotting graphs

4 LOADING DATASETS

There are two datasets which are used are:

'application_data.csv' contains all the information of the client at the time of application.
The data is about whether a **client has payment difficulties.**

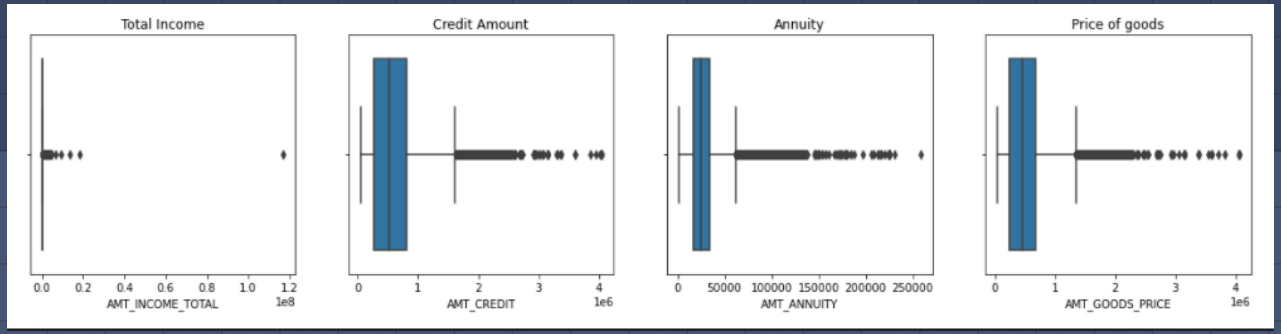
'previous_application.csv' contains information about the client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**



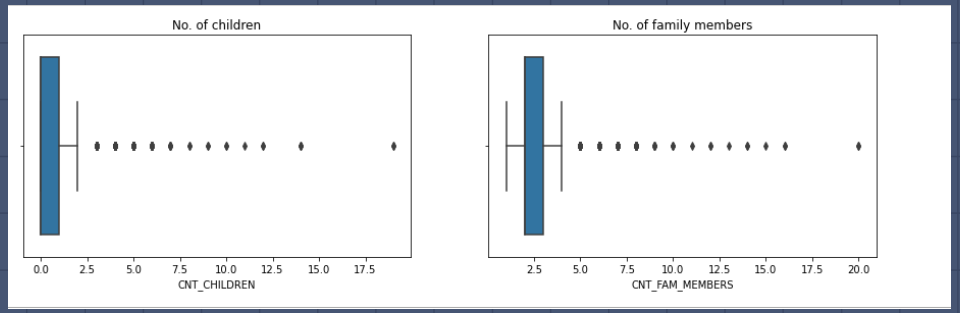
5 DATA INSPECTION AND CLEANING- Application Data

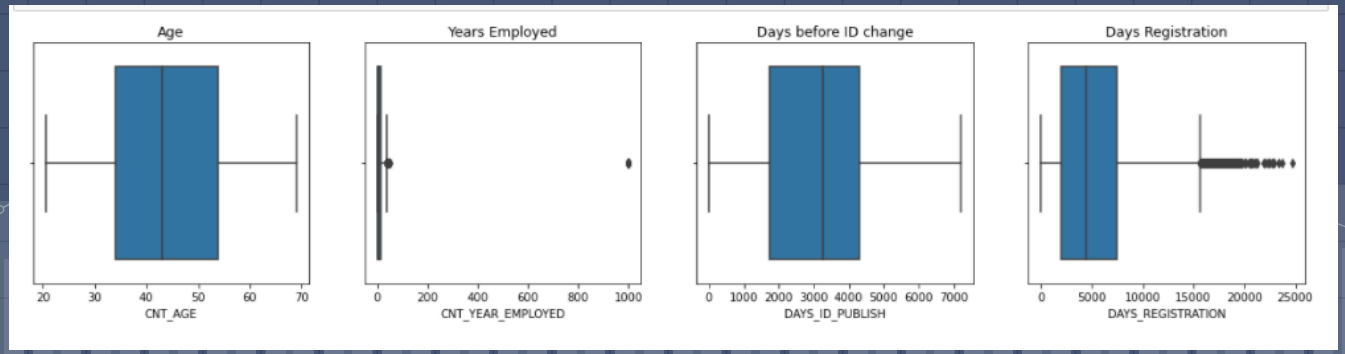
1) Checked the number of rows(307511) and columns(122) of data.
2) Checked the data types of all columns.
3) Searched for missing values in each column.
4) Dropped all the columns having missing values >45%, which deleted 49 columns and the new shape of the data is (305711,73).
5) Imputed values for 5 columns: OCCUPATION_TYPE, AMT_GOODS_PRICE, NAME_TYPE_SUITE, CNT_FAM_MEMBERS.
6) Dropped more columns which were of no use to the target variable leaving us with the final shape of (307511,31).
7) Checked invalid entries in columns and replaced/corrected them if any.
8) Binned two continuous variables into categorical columns.

HANDLING OUTLIERS



Checked for outliers in numerical columns and imputed them for 5 variables by capping.

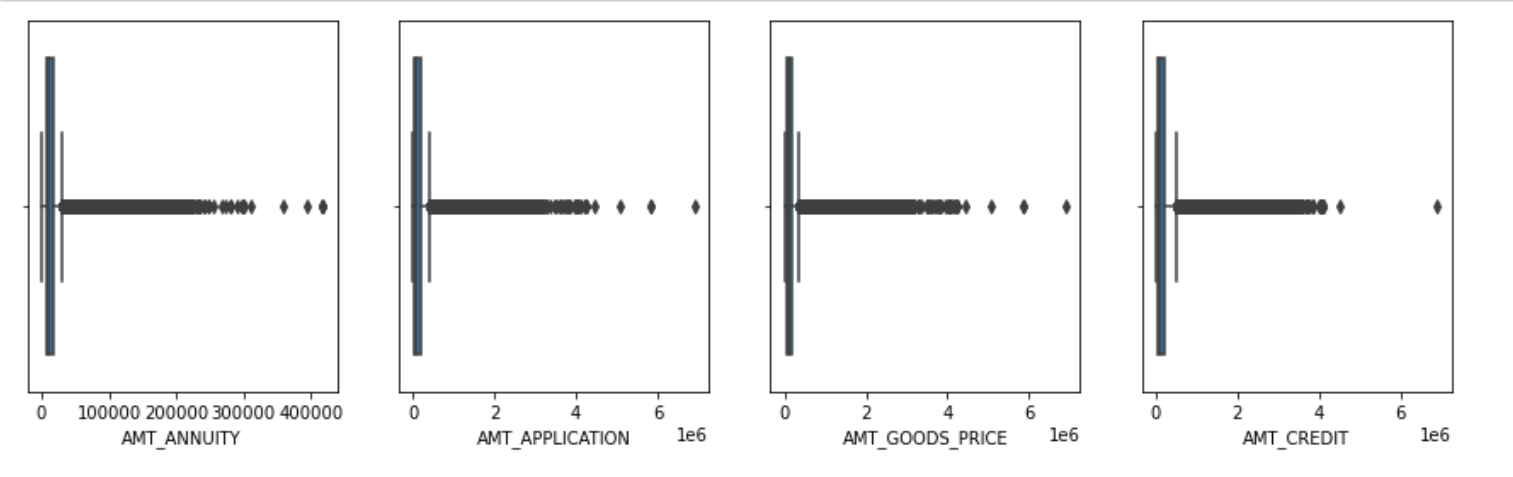




DATA INSPECTION AND CLEANING-Previous Data

1) Checked the number of rows(1670214) and columns(37) of data.
2) Checked the data types of all columns.
3) Searched for missing values in each column.
4) Dropped all the columns having missing values >45%, which deleted 5 columns and the new shape of the data is (1670214,32).
5) Imputed values for 5 columns: AMT_GOODS_PRICE, AMT_ANNUITY, AMT_CREDIT, CNT_PAYMENT, PRODUCT_COMBINATION.
6) Checked invalid entries in columns and replaced/corrected them if any.

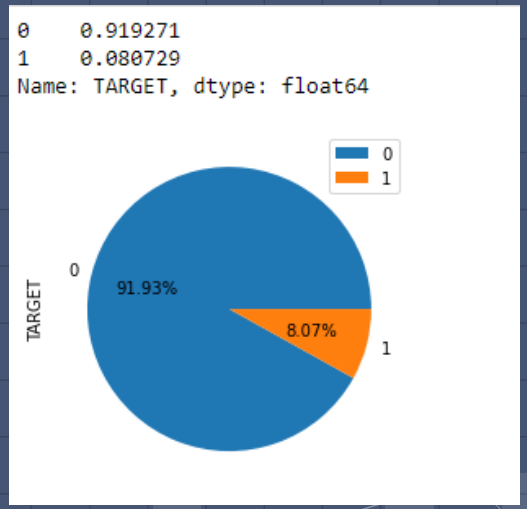
HANDLING OUTLIERS



Checked for outliers in numerical columns and imputed them for 4 variables by capping.

ANALYSIS- UNIVARIATE

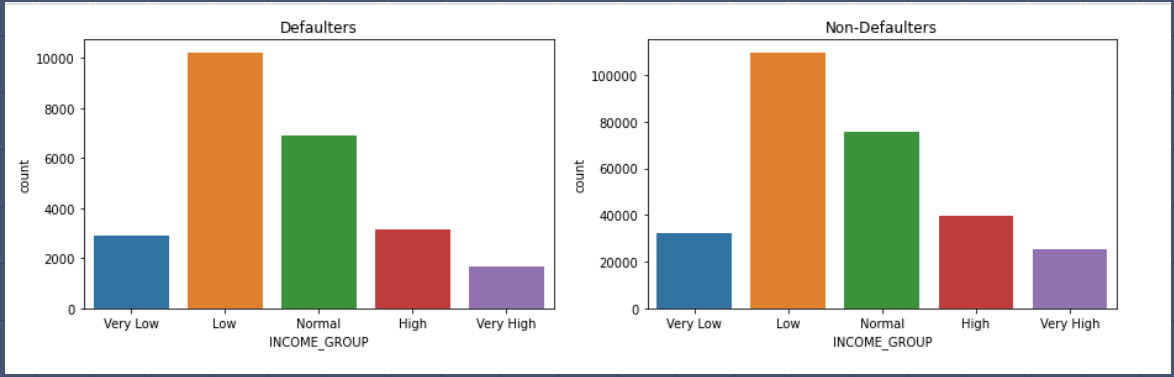
IMBALANCE PERCENTAGE



There are 91.93% people who have paid the loan amount and 8.07% of people who have not paid the loan and are defaulters.

ANALYSIS- UNIVARIATE

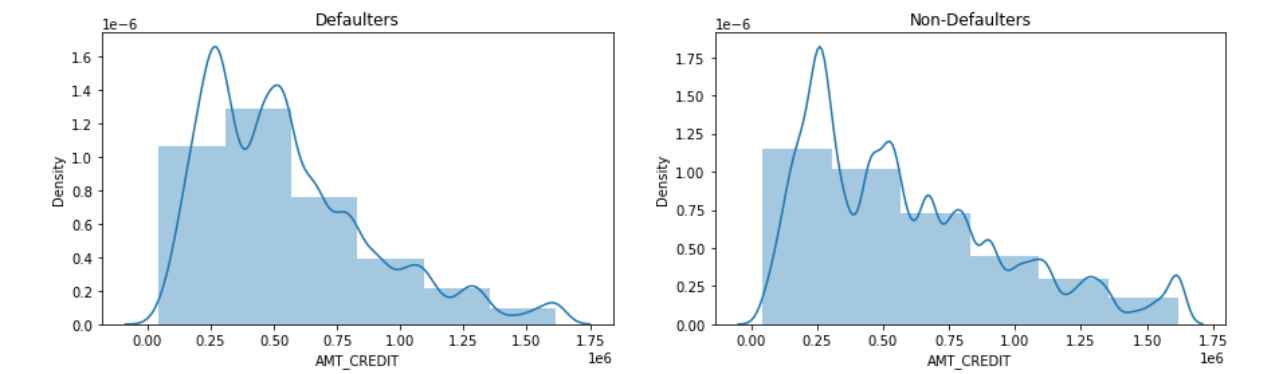
INCOME GROUP VS TARGET



From the above graphs it is clearly visible that people with low income tend to have highest number in defaulters as well as non-defaulters.

ANALYSIS- UNIVARIATE

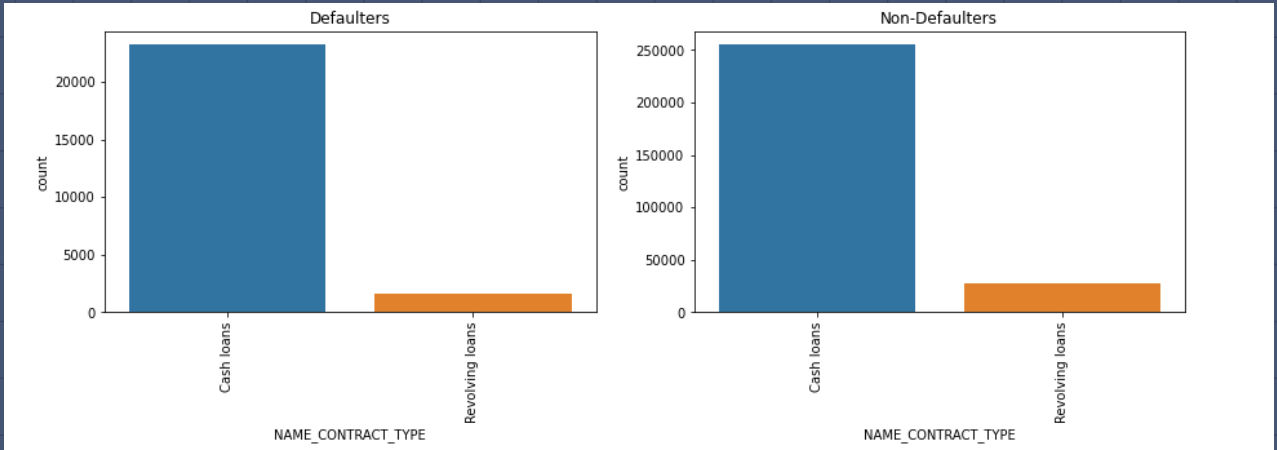
CREDIT AMOUNT VS TARGET



Non-defaulters seems to follow a general trend where most people take a loan of around 3 lakhs and as the loan amount increases, number of people decreases because not all people take huge loans. In defaulters most people seem to have taken a loan of 5 lakhs.

ANALYSIS- UNIVARIATE

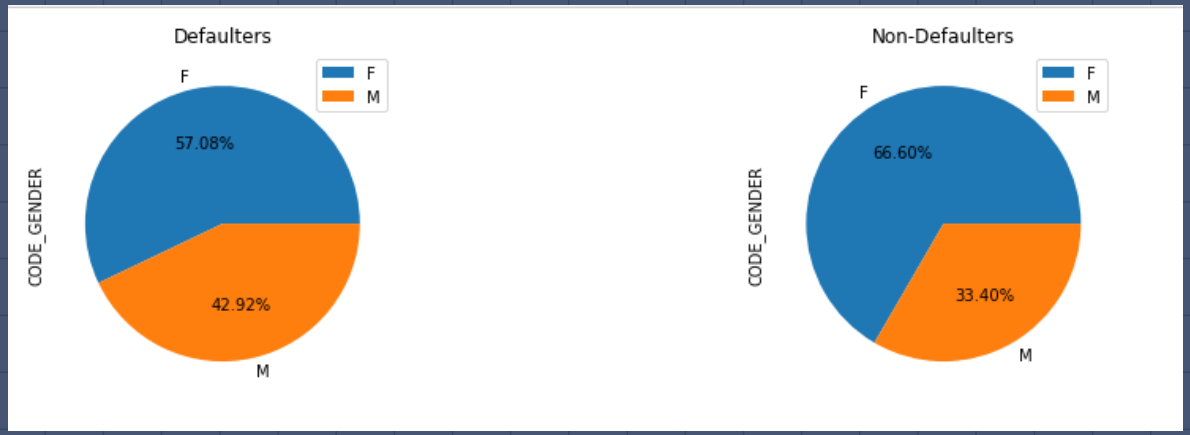
CONTRACT TYPE VS TARGET



Maximum defaulters as well as non-defaulters have taken cash loans and a very small amount of people have taken revolving loans.

ANALYSIS- UNIVARIATE

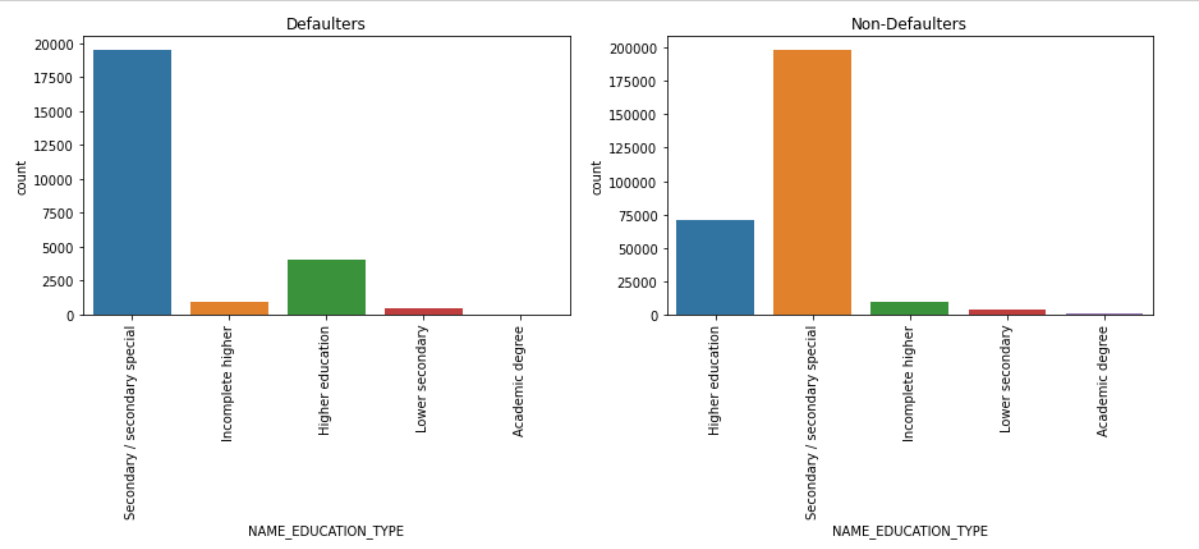
GENDER VS TARGET



Females are maximum in both defaulters and non-defaulters

ANALYSIS- UNIVARIATE

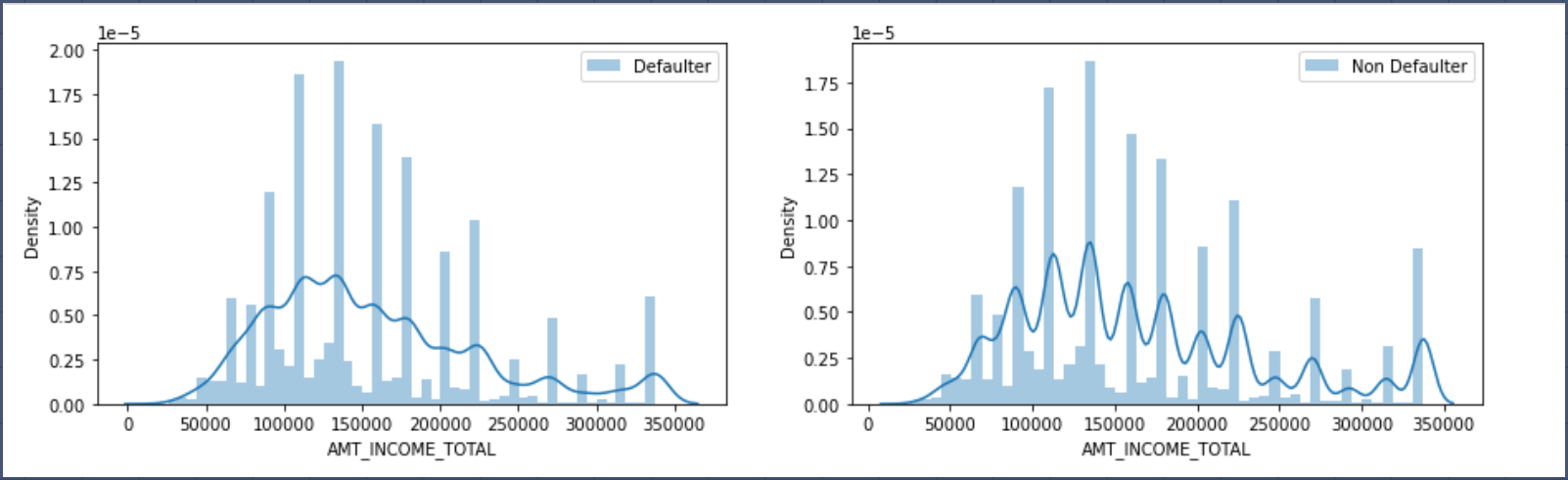
EDUCATION TYPE VS TARGET



People with secondary/secondary special education are maximum in number in both defaulters and non -defaulters, followed by people with higher education.

ANALYSIS- UNIVARIATE

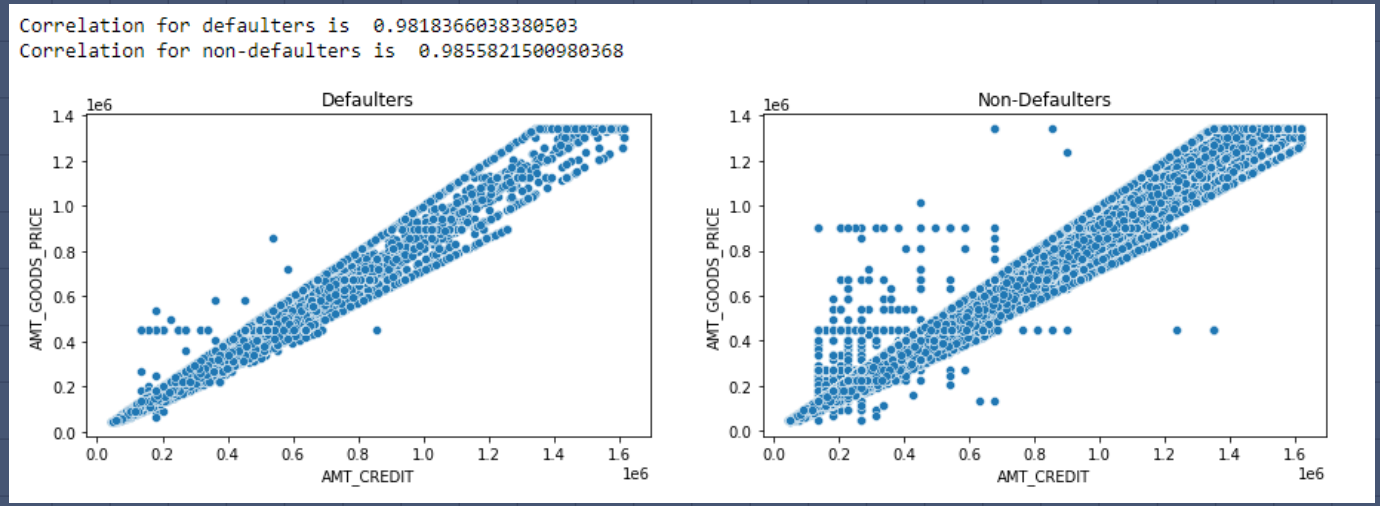
GENDER VS TARGET



In defaulters, people with the income between 1L-1.5L are more in number than people with the income beyond 2.5L. This depicts that higher income people are still less defaulting as compared to lower income people.
In non-defaulters we can see a normal high-low trend.

ANALYSIS- BIVARIATE

GOODS PRICE VS CREDIT

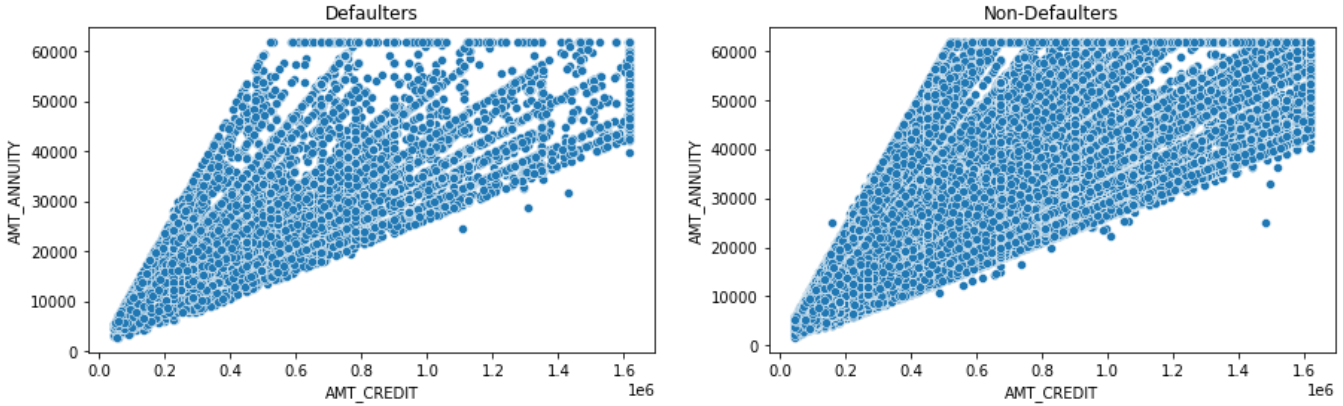


From the above plot it can be seen that credit amount is highly correlated with the price of goods in both defaulters and non-defaulters. This is also verified by the value of correlation coefficient which is 0.98. This seems to be fairly true that more the price of goods more is the amount of loan which is taken.

ANALYSIS- BIVARIATE

ANNUITY VS CREDIT

```
Correlation for defaulters is  0.7601234136216812
Correlation for non-defaulters is  0.7948078743243145
```

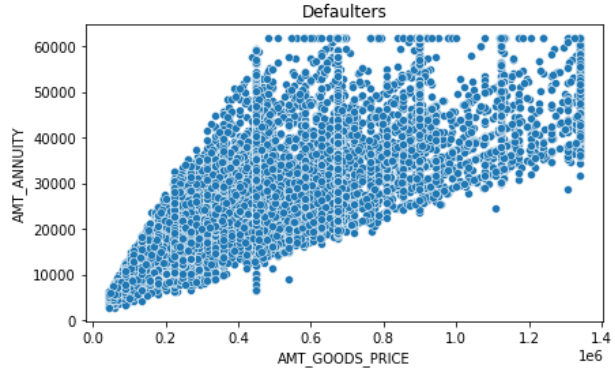


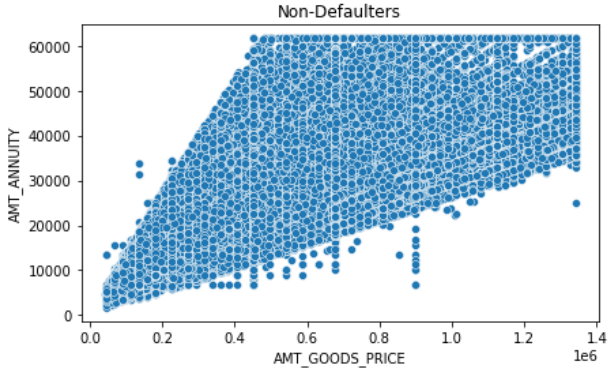
There seems to be a normal correlation between the annuity amount and credit amount.

ANALYSIS- BIVARIATE

GOODS PRICE VS CREDIT

```
Correlation for defaulters is  0.7602866472620412
Correlation for non-defaulters is  0.7973154338305388
```



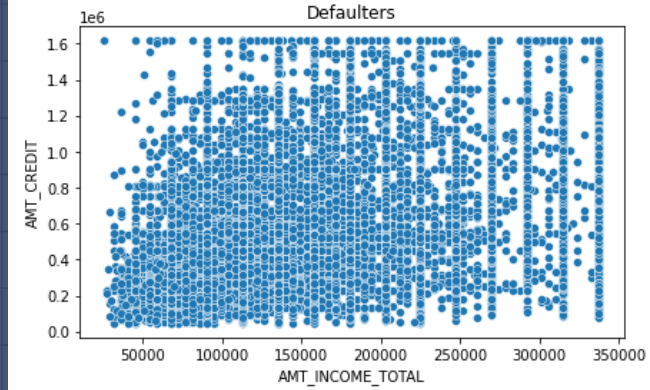


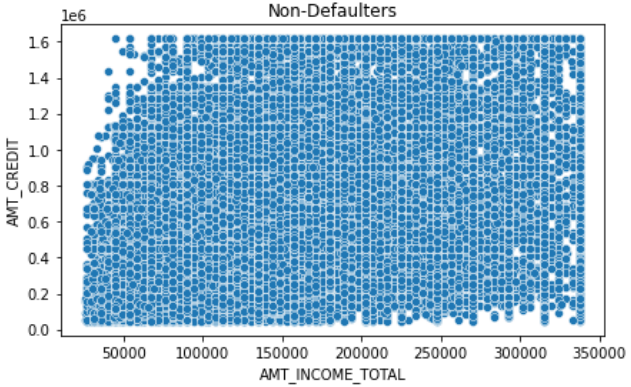
There seems to be a normal correlation between price of goods and the amount of annuity but the causation between the two is not guaranteed.

ANALYSIS- BIVARIATE

TOTAL INCOME VS CREDIT

```
Correlation for defaulters is  0.356198773589607
Correlation for non-defaulters is  0.41430852924965916
```

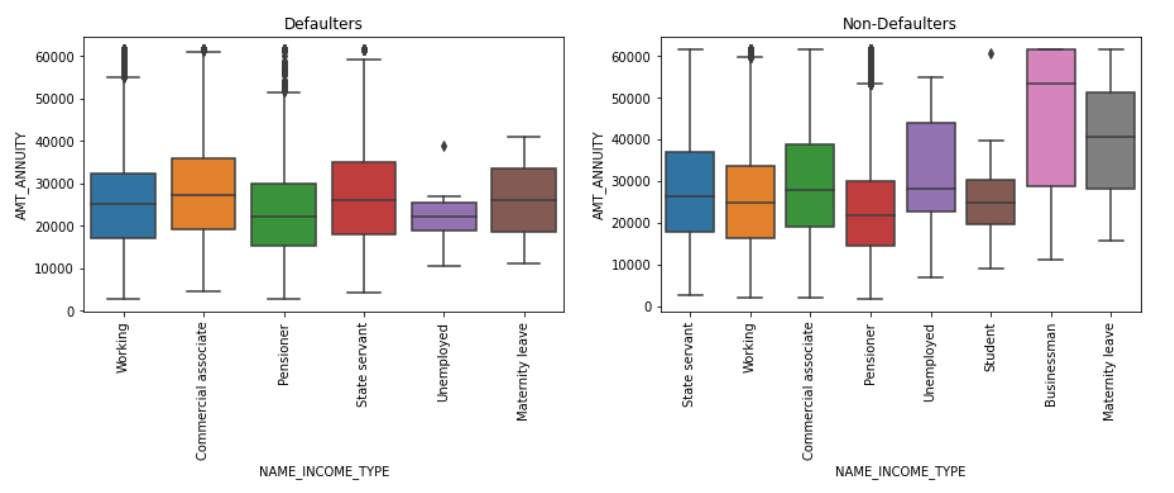




The correlation between income of client and the amount of loan does not seems to have a very good correlation for defaulters and non-defaulters both. It is expected that people with higher income are able to take huge loan amount so that they can repay it easily but this is also true that people with low income also take huge amount loan to fulfill their luxurious needs. This can be seen from the defaulters graph that the points are not very dense where the loan amount and credit amount increases.

ANALYSIS- BIVARIATE

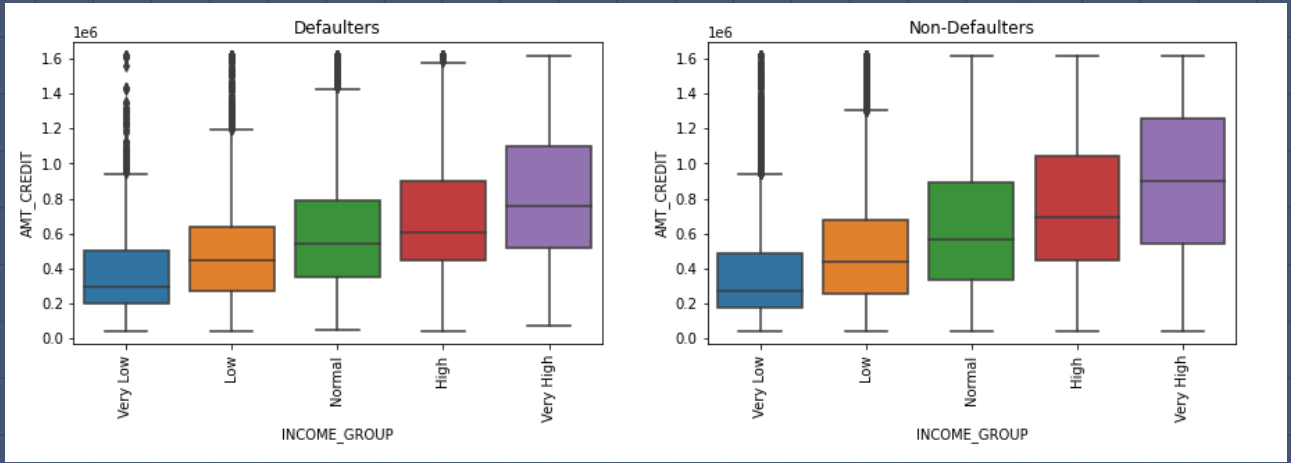
INCOME TYPE VS ANNUITY



From the above plot it can be seen that:
In the defaulters it can be seen that all people have median annuity amount between 20,000 - 30,000.
Also there are no businessmen and students who possess an annuity amount in this category.
In the non-defaulters businessmen have the highest median annuity amount followed by people on maternity leave.

ANALYSIS- BIVARIATE

INCOME GROUP VS CREDIT

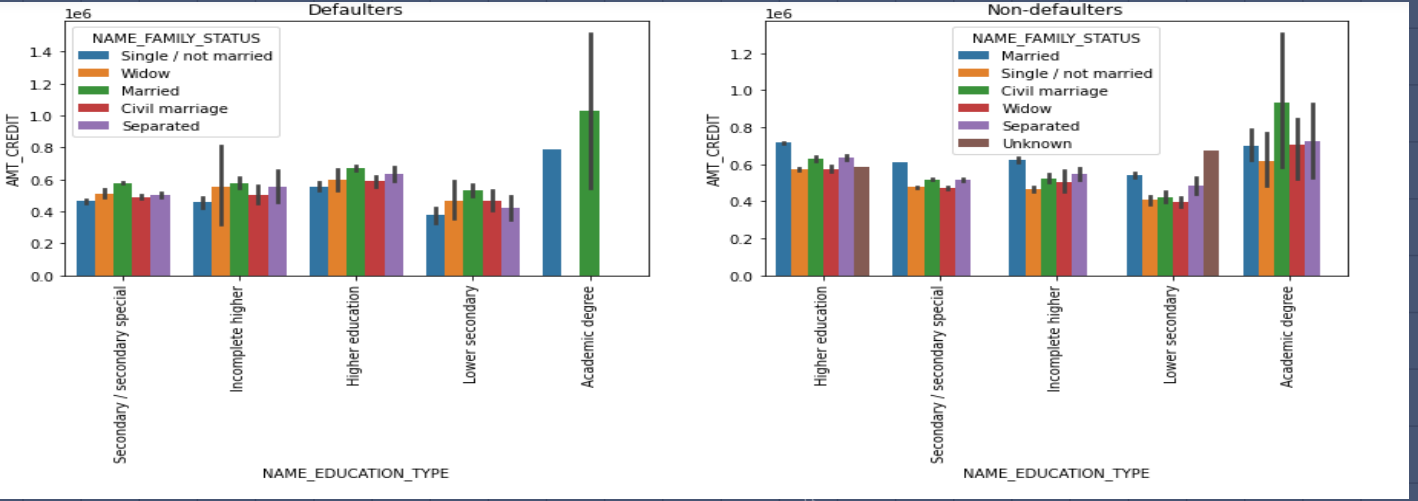


Almost similar trend is seen in both the graphs. In the defaulters rich people have taken a median loan amount of 7L whereas in non-defaulters the median loan amount for the rich section is almost 10L. This is quite evident as bank has build trust on non-defaulters and have given a credit.

Also a normal trend is visible where as the income is increasing, the median amount of loan is also increasing.

ANALYSIS- BIVARIATE

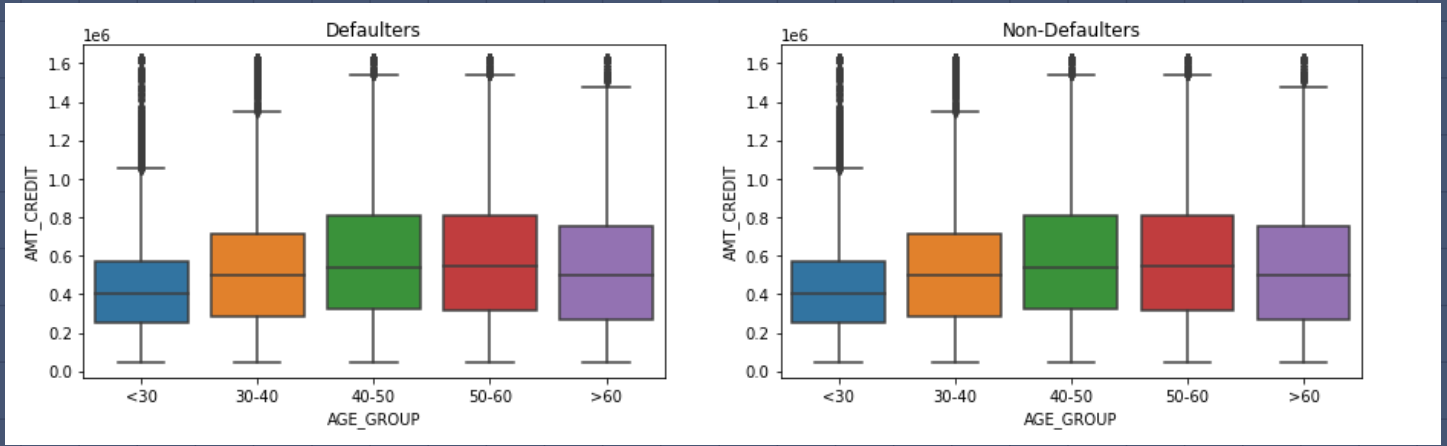
EDUCATION TYPE VS CREDIT



There are no widows, civil marriage people and separated people having an academic degree and have failed in repaying the loan.
In the defaulters, all the married people having any education degree has taken maximum credit. Same goes with the non-defaulters except for people having Academic degree

ANALYSIS- BIVARIATE

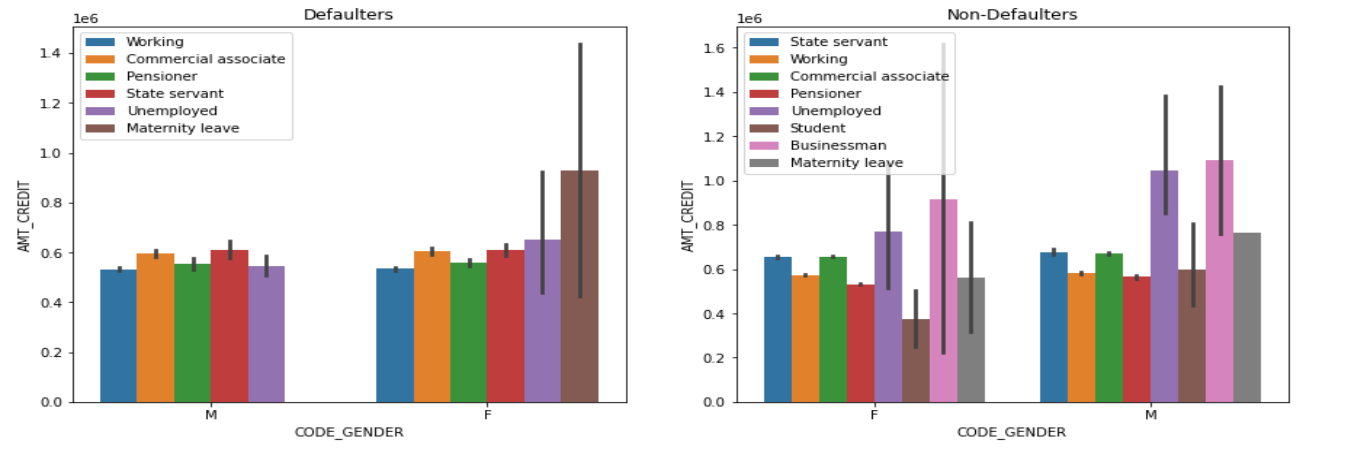
AGE GROUP VS CREDIT



In both cases maximum credit is taken by the people in age group of 40-50, followed by people in the age group of 50-60. In defaulters people of age less than 30 years have low credit amount but still they are unable to pay the loan. This could be because these people still don't have stable jobs in hand.

ANALYSIS- BIVARIATE

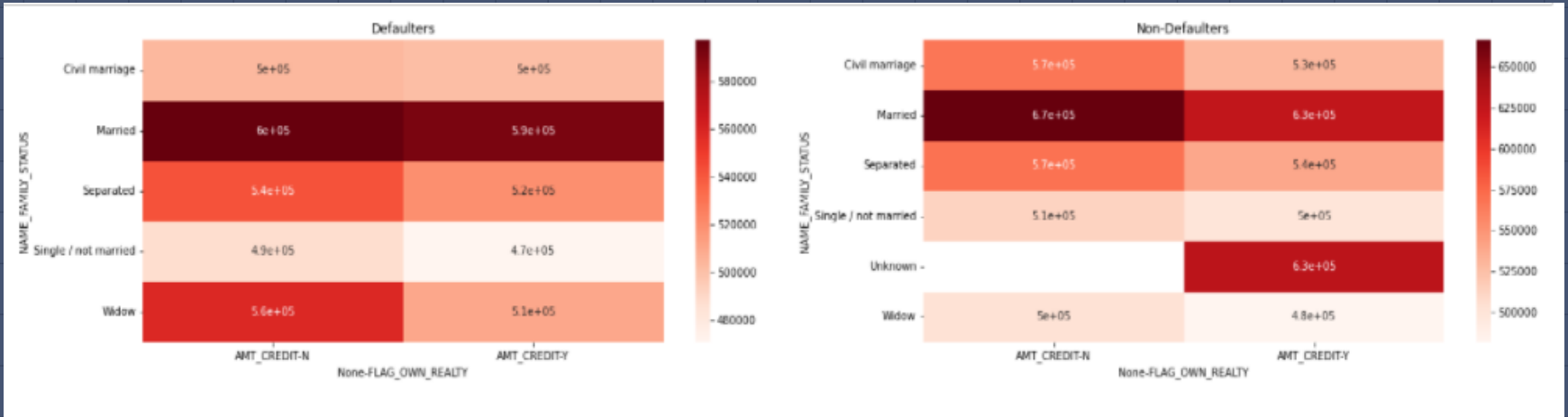
GENDER VS CREDIT



Pregnant females tend to take high credit and fail in repaying whereas pregnant females who have taken low credit have repaid successfully.
There are no students and businessmen who have failed to replay the loan
An unusual trend can be seen where males and females who are unemployed have managed to repay the loan and have also taken huge credit.

ANALYSIS- MULTIVARIATE

REALTY VS FAMILY STATUS VS CREDIT



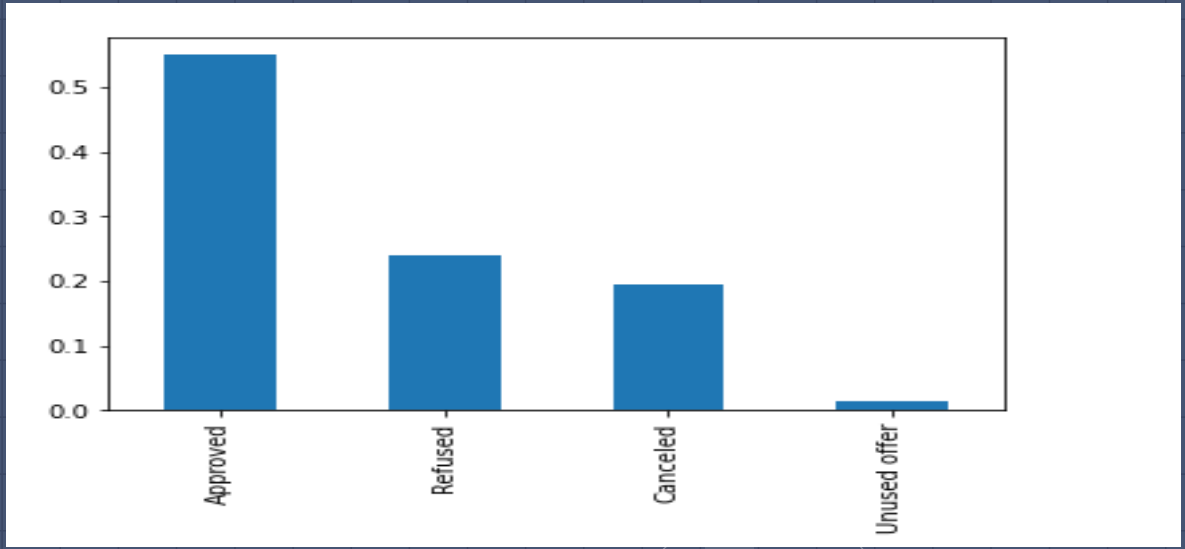
Married people who does not have a realty tend to take huge loans.
Widows who does not have realty have taken huge loan and hence are unable to repay but there are some widows who do not have realty but have repaid the loan amount.
Civil marriage people who are defaulters are having same amount of credit which is less than those who are non-defaulters but are still unable to repay the loan.

ANALYSIS- Using previous data

The current data and previous data has a common column --> 'Current ID'. Let's merge these two data on this column and check if we can get some good insights.

ANALYSIS

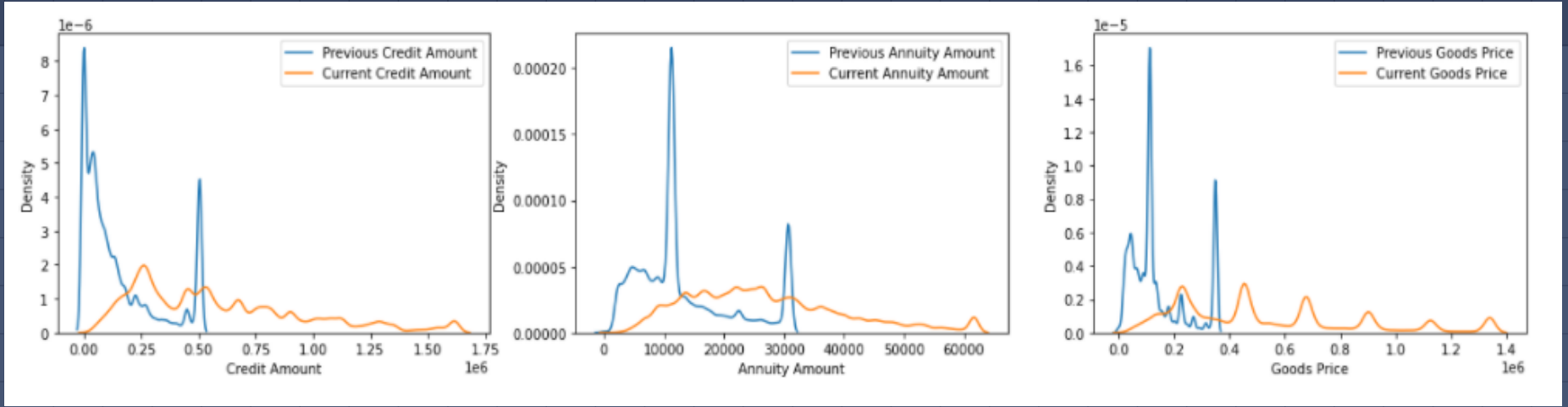
CONTRACT STATUS OF DEFAULTERS



It can be seen that amongst the defaulters 25% of people's applications were rejected in their previous applications and almost 50% applications were approved.

ANALYSIS

ANNUITY AMOUNT, CREDIT AMOUNT, GOODS PRICE OF PREVIOUS AND CURRENT APPLICATIONS



Insights from the graph:

The previous applicants are maximum in the credit range of 1L-2L, but a sudden downfall is seen for the credit range of 2L-4L with again a rise at 5L. The current applicants seems to follow a normal trend.
An almost similar trend is seen in the price of goods graph for previous applications, but in new applications maximum people have purchased goods of around 3L, 5L, 7L, i.e. in odd amounts.
All the graphs clearly depicts that the maximums of each variable in previous applications are less than the maximums of each variable in the current applications.

SUMMARY

- In current applications cash loans were maximum whereas in previous applications consumer loans were maximum.
- Businessmen and students are not amongst the defaulters therefore banks can trust them in giving loans.
- Women are more liable to default than men.
- Banks can give loans to pensioners.
- Goods price and credit show a very high and positive correlation.
- Annuity and goods price are positively correlated but there is no causation between the two.
- Maximums of most variables in previous applications are less than the maximums of most variables in the current applications.