CREDIT EDA–
CASE STUDY

Understanding how Consumer Attributes and Loan Attributes influence the tendency of default

10th JUNE2019

STEPS FOLLOWED FOR THE CREDIT EDA CASE STUDY

- Data Preparation and Exploration
- Data Cleaning
- Data Analysis
- Data Visualisation
- Inferences and Recommendations

➢ The stats and the shape of the datasets where well observed initially before any action.
➢ All the missing values and irrelevant columns were treated(either imputed or dropped) accordingly.
➢ Operations where done on each and every important column and when necessary existing data frame was splitted on to relevant datasets.
➢ Utilised different plots offered by Python according to varaiable/s in analysis.
➢ Analysis, Anomalies, Inferences were listed accordingly.

DATASETS USED

- Application Dataset -> Contains all the information of the client at the time of application.

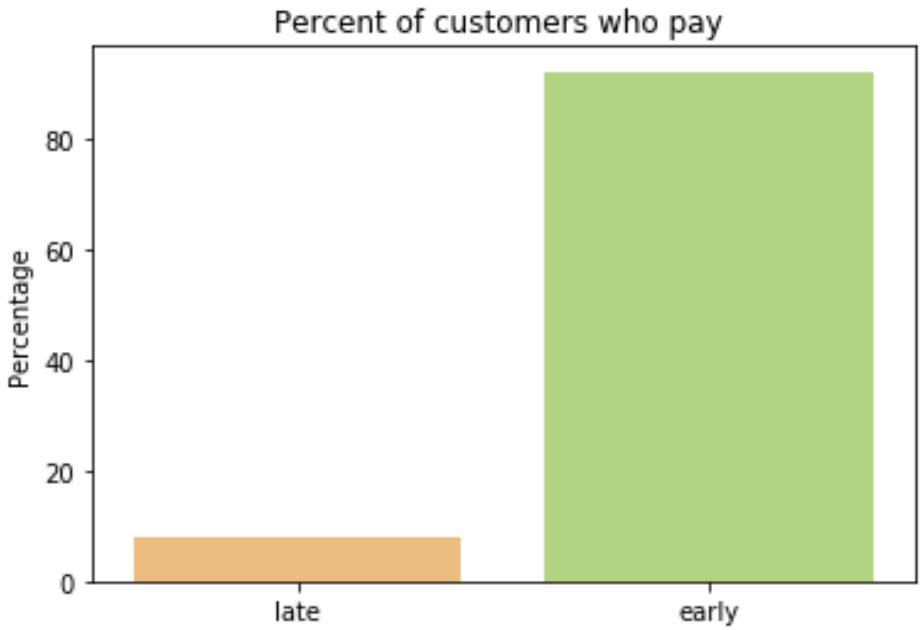
The data is about whether a **client has payment difficulties.**

- Previous Application Dataset -> 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.**

WHAT APPLICATION DATA SHOWS US?

Percentage of Applicants who pay Late vs Early.



	Percentage
late	7.724696
early	92.275304

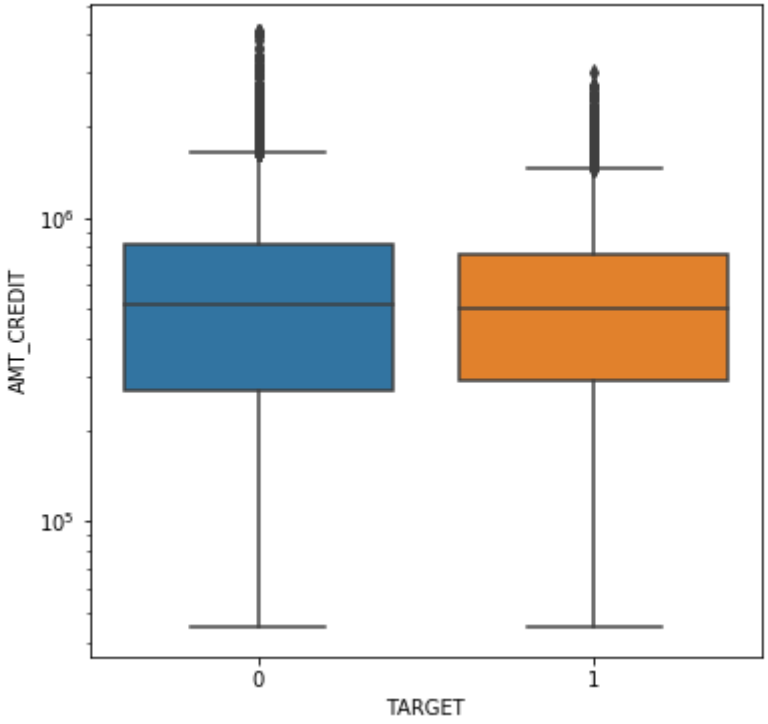
➢ So among the given list of applicants around 7.72% of the applicants are the one with payment difficulties or in other sense had late payment instances in the past. While 92% of the applicants are the one with good payment history and hence an asset to the bank.

Outlier Detection

➢ Here from the above boxplot between the Target and the Income of the client, we can find, the income for the clients who are having payment difficulties and other cases both are in similiar range with median value almost same.

➢ But most importantly, what we can notice is outlier values present in the dataset. There are outlier values present for both the Target cases, **but we can easily spot a High Income client present who had previously a payment difficulty**(Target Variable 0).

.



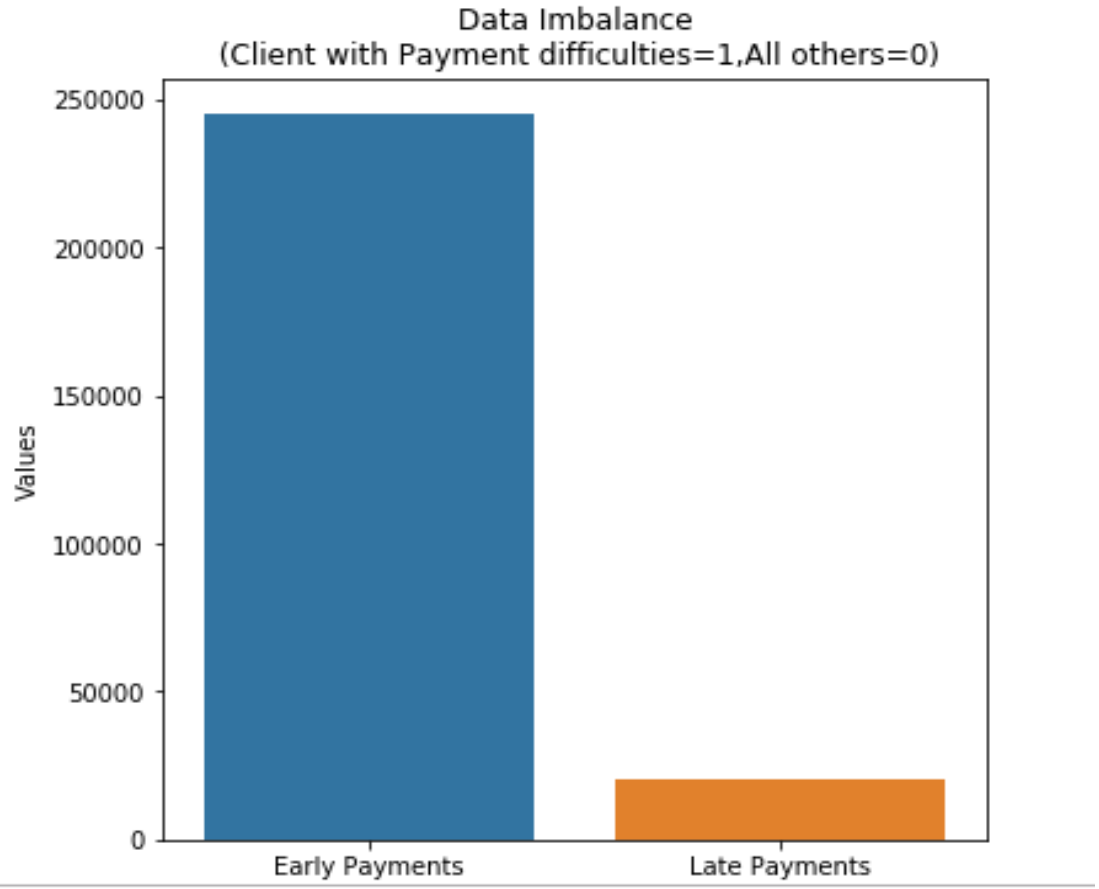
Data Imbalance for the Target Variable

	Values
Early Payments	245014
Late Payments	20511

- We can find a data imbalance here with respect to the Target variable of the dataset since the data points for Target=1 i.e applicants with payment difficulties is at around 7.72% only compared to the total number of applicants.
- This High Imbalance of Target Data can be removed with random re-sampling of the collected sample data in order to avoid any bias.

- ***The ratio of 12:1 can create bias in data and thus is important to remove or redistribute in sampling data as uniformly distributed random data***

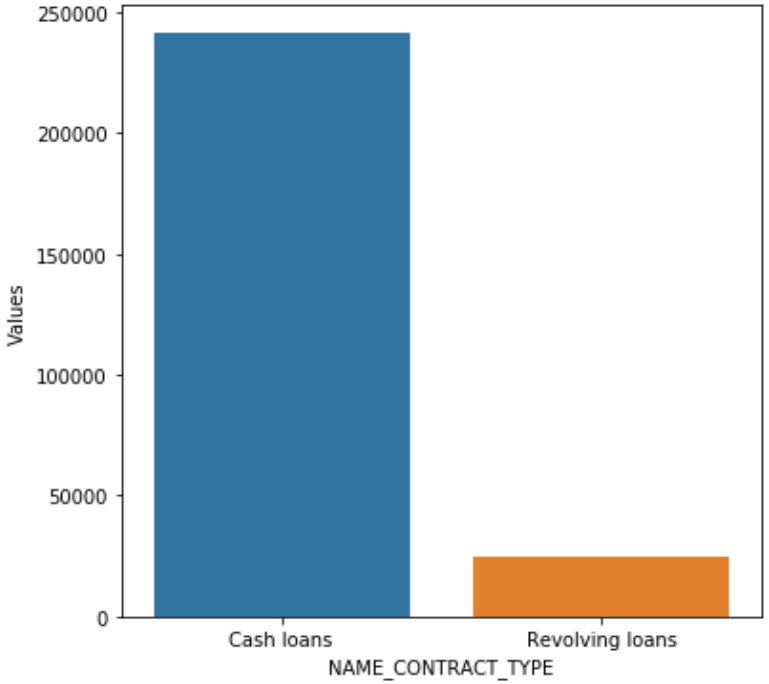
.



Amount Credit

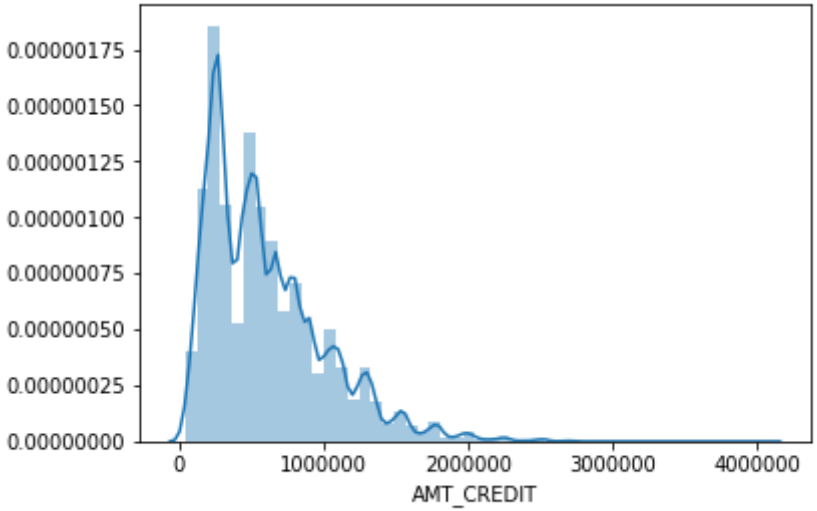
➢ Density Plot for Credited amount of the loan for the clients shows that majority of the loan amounts of the clients are below 10 Lakhs.

.



➢ **This also improves the credit Scores of Customers and helps banks to eliminate defaulters with low Credit Scores.**

.



Cash Loans vs Revolving Loans

➢**Around 90% of the client Loans are Cash Loans. Rest 10% are the only revolving Loans.**

➢We find that the Number of People who don't have monetary problems are higher than the ones who have monetary Problems.
But, unfortunately the Number of Cash Loans given out to Customers are higher than Revolving Loans. Revolving Loans has few benefits for both customers as well as Banks.

➢***For Customers:*** Customers who do pay loans on time and are capable, do not need to apply loan separately each time. This creates ease for customers to apply loans and avail easy loans without hassle, thus increasing chances of Customer retention.

➢***For Banks:*** Banks can generate more revenue and easily charge more interest for successive loans applied from same customer and convince customer to apply loan from them without filing for a fresh loan anywhere else.

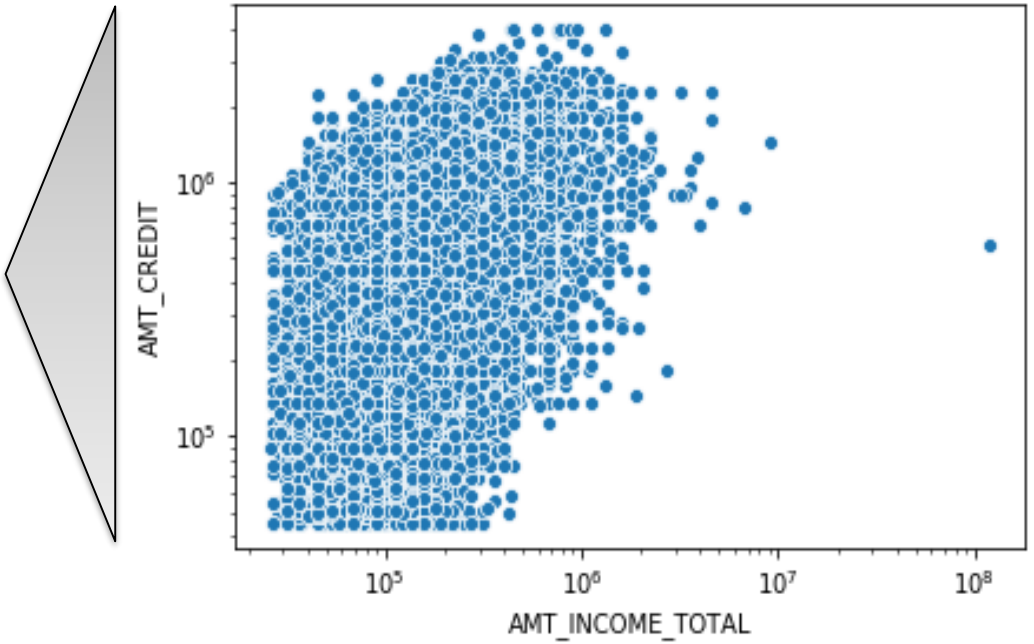
Credited Amount and Total Income Amount Related?

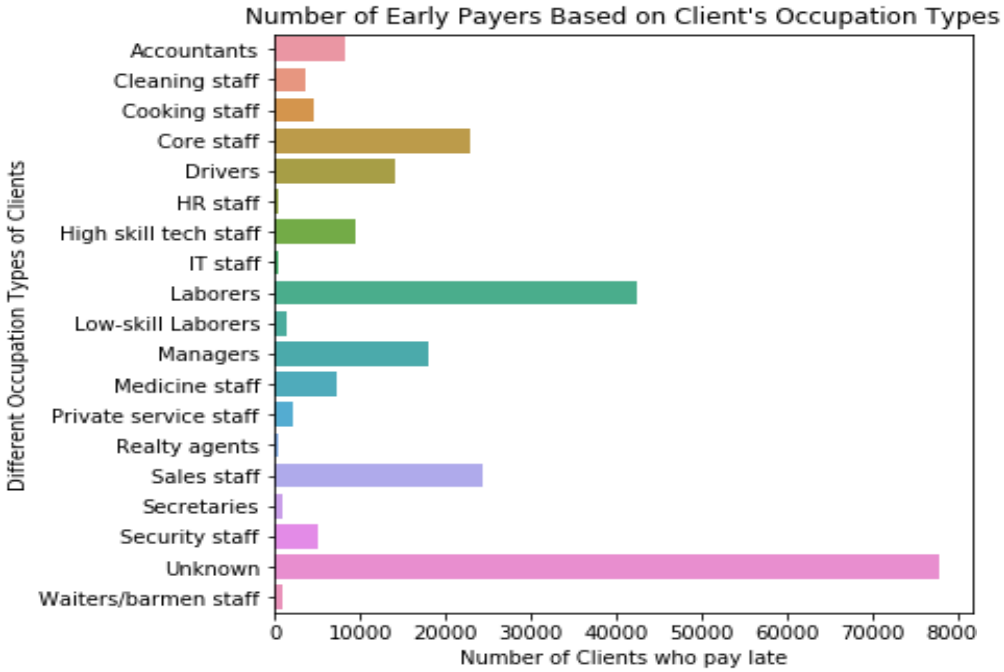
➢ Heat map here shows very low correlation between the Income of the Applicants and the Loan credited amount which in turn implies that even less income applicants have opted out for a large Loan amount and also high income applicant would have opted out for Small Loan amounts.

.

➢ This scatter plot gives us an important insight **that the amount of default credit given to the people with lower total income is quite higher which may lead to defaulting.**
Also, above total income greater than 1000000,the amount of credit decreases. This can be due to be reasons like
1.High income persons turn to default likely
2.The banks may not have sufficient funds to credit customers.

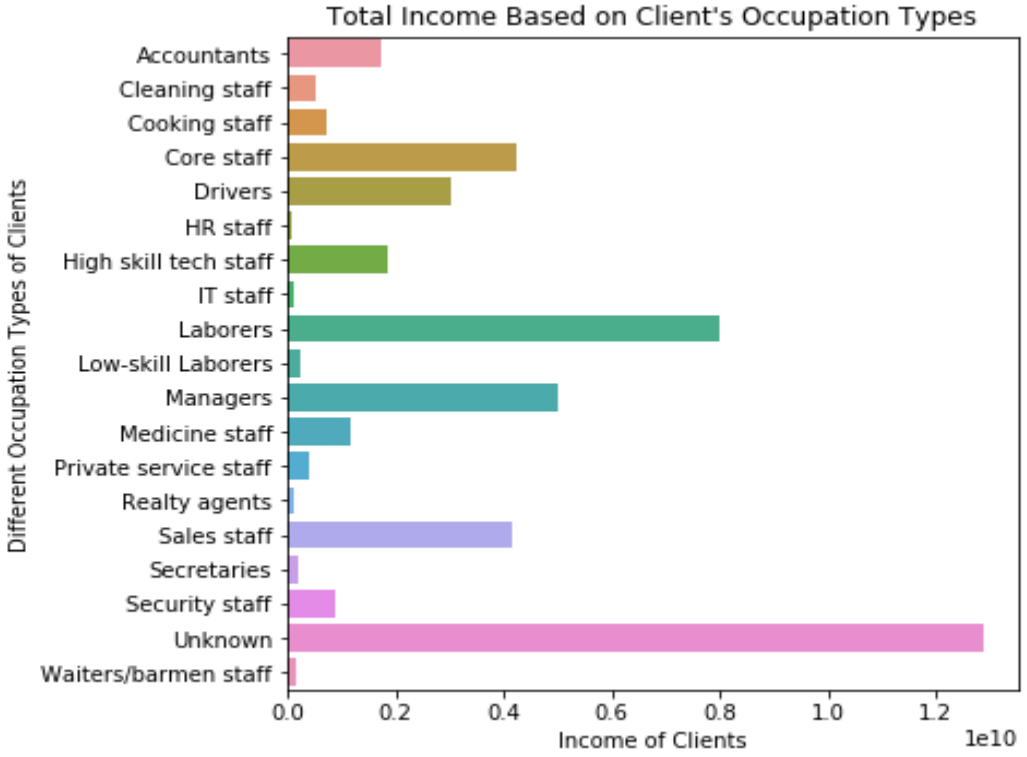
.



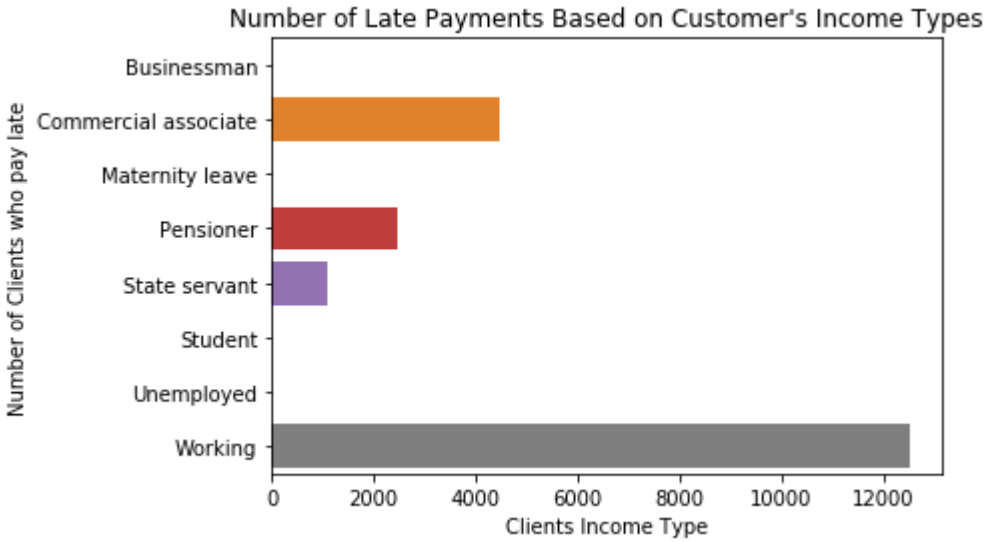


Clients Occupation Types and the Number of Early Payers

Clients Occupation Types and the Income of the Clients



Customer’s Income Types and the Number of Late Payments



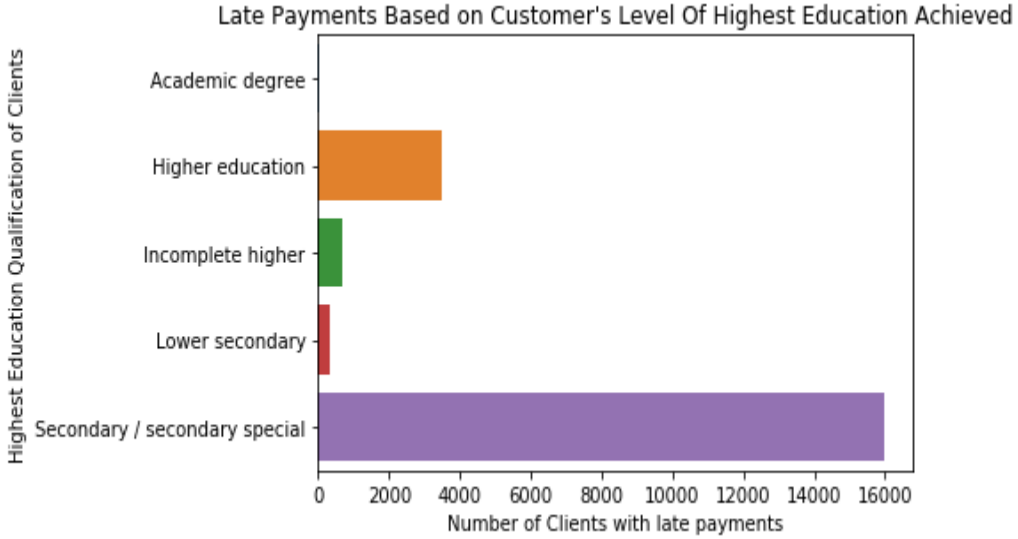
➢ Working Class of Customers find it hard to pay off their debts on time. But, on the other hand it is also seen that Businessman do not pay late.

But very suspicious insights obtained is that how Unemployed Customers and Students do not delay in paying debts.

Customers with Maternity leaves can still get their maternity benefits with their regular salary..

.

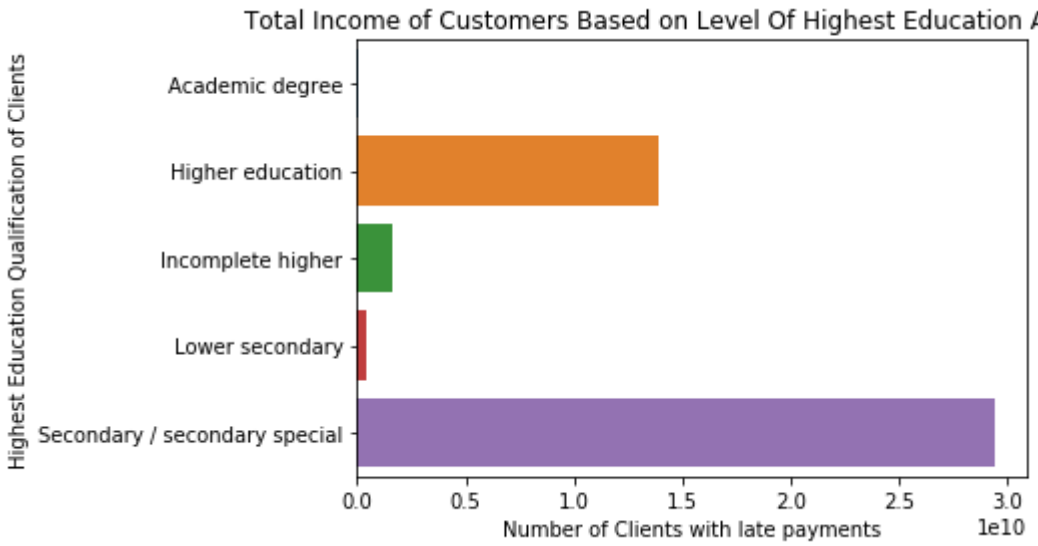
Late Payments and Customer’s Level of Highest Education



➢ Applicants with Secondary/Secondary special level of Education as their highest education level are the highest Late Payers.

.

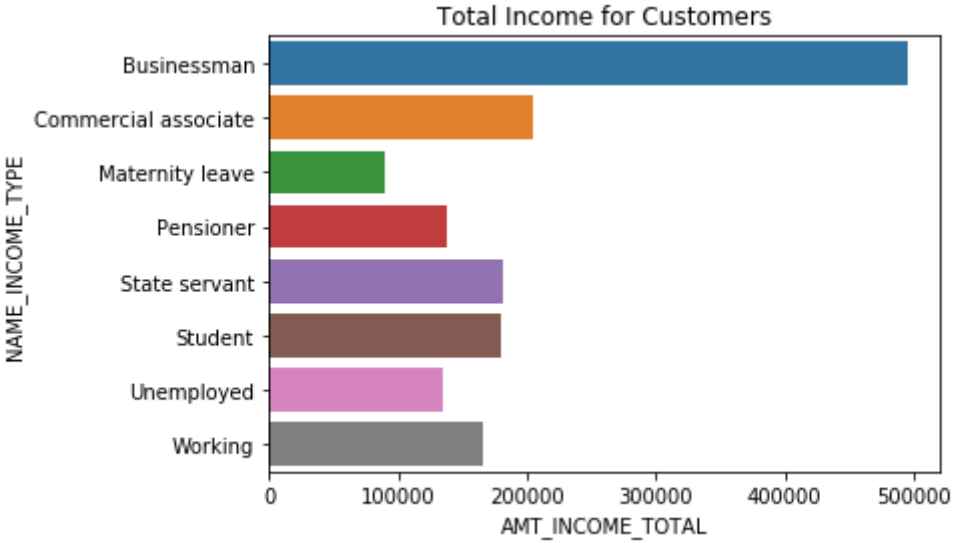
Anomaly for the Lower Secondary Education Level Applicants



➢ **Anomaly**: One Strange thing about the above analysis, how can the people with the lowest Education Recieved,i.e Lower Secondary has the minimum times where they paid off their debts late, even though their total income is very less

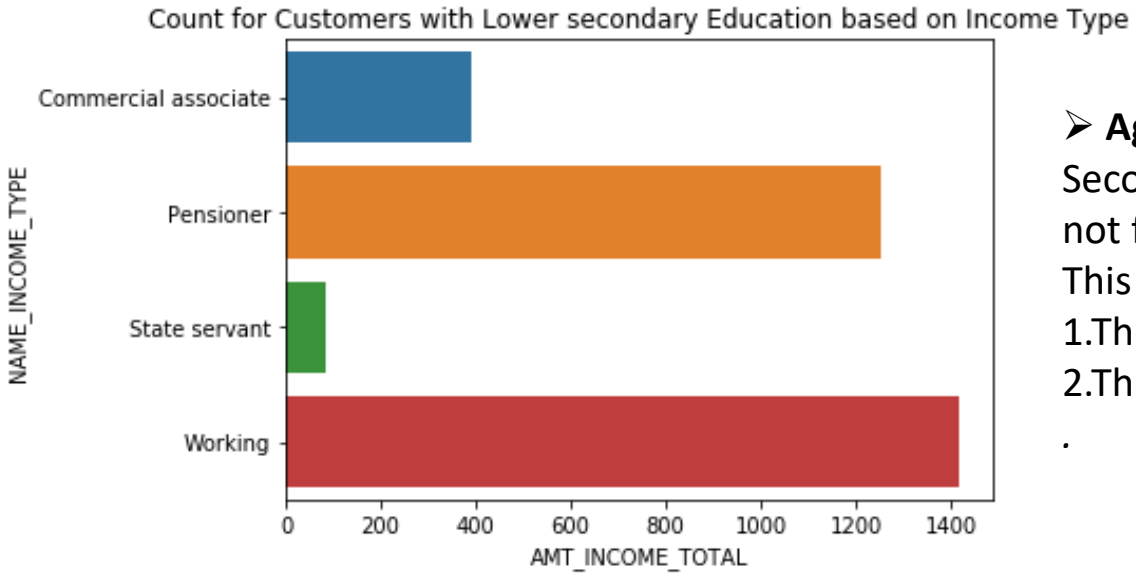
.

Income comparison for the Applicants



➢ Out of all the applicants, Businessmen are the Highest Income applicants followed by Commercial associate.

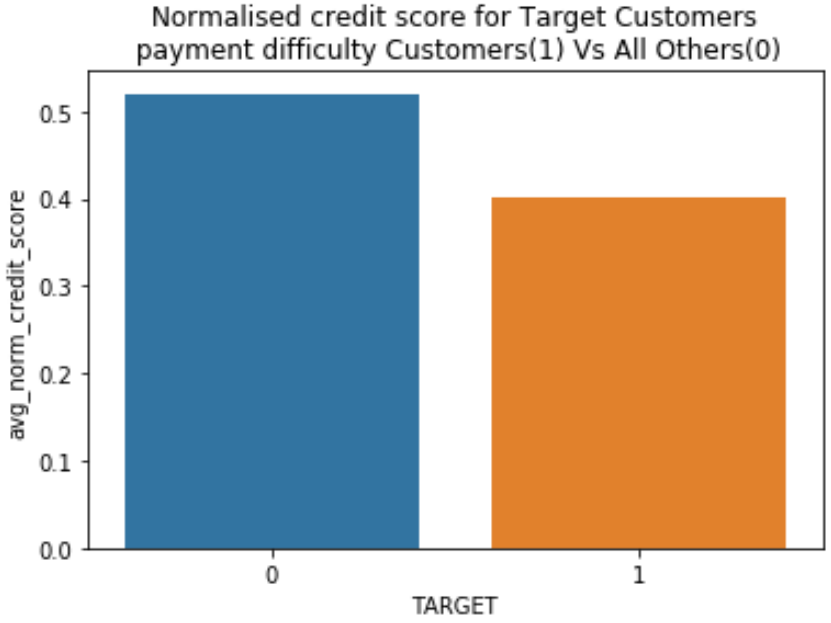
Lower Secondary Education Level Applicants Income Types?



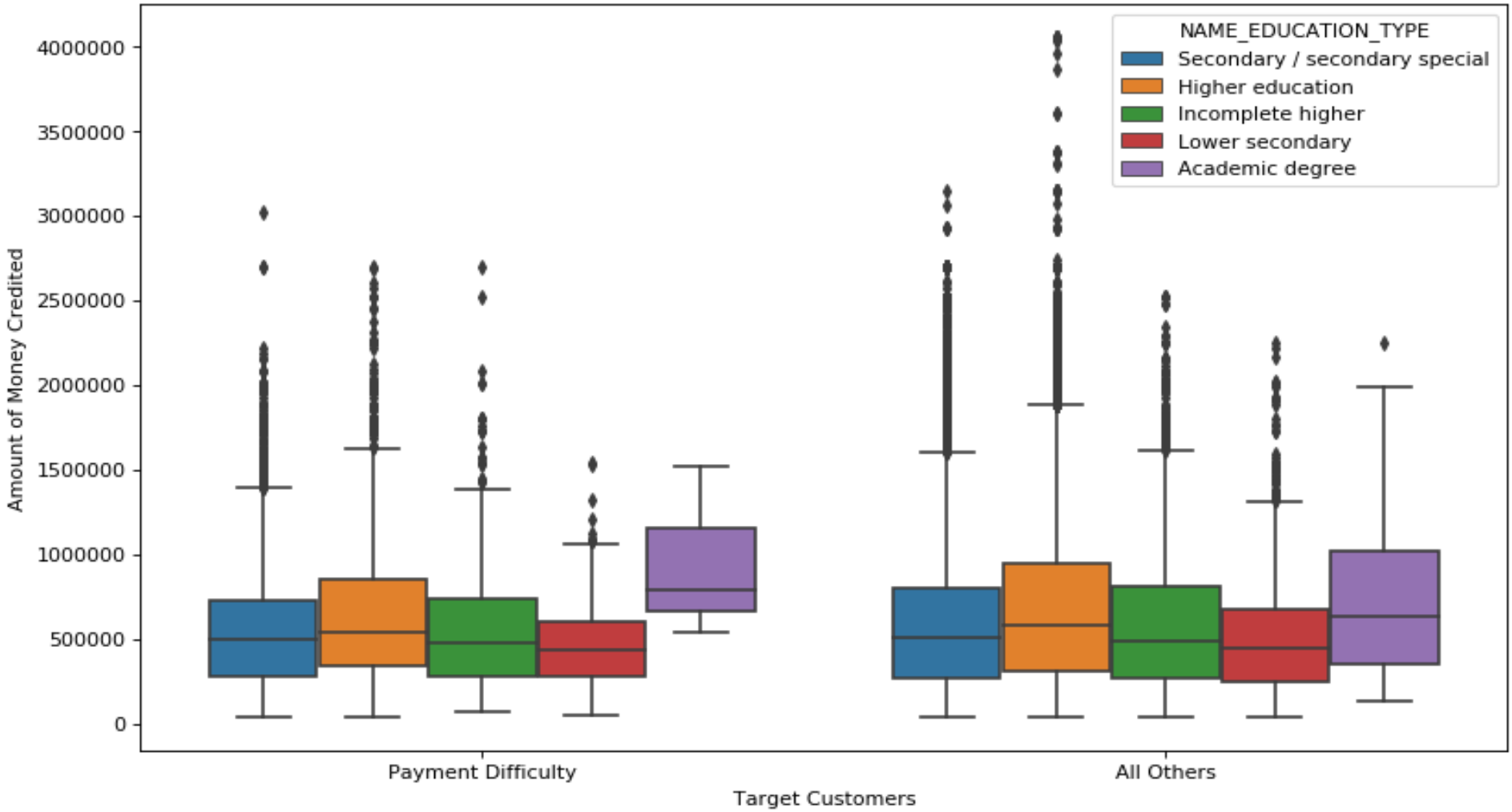
➢ **Again Anomaly**: We see that Customers with Lower Secondary Education are mostly from Working Class, but still do not face any difficulty in paying their debts.
This can be due to 2 reasons:
1.They have a hidden source of income which is not mentioned.
2.They data may be corrupt or wrong data is provided.
.

Normalised CREDIT SCORE

➢ The Average Normalised Credit Score of Customers who pay on time is **22.83 %** higher than those who pay late due to payment difficulties.

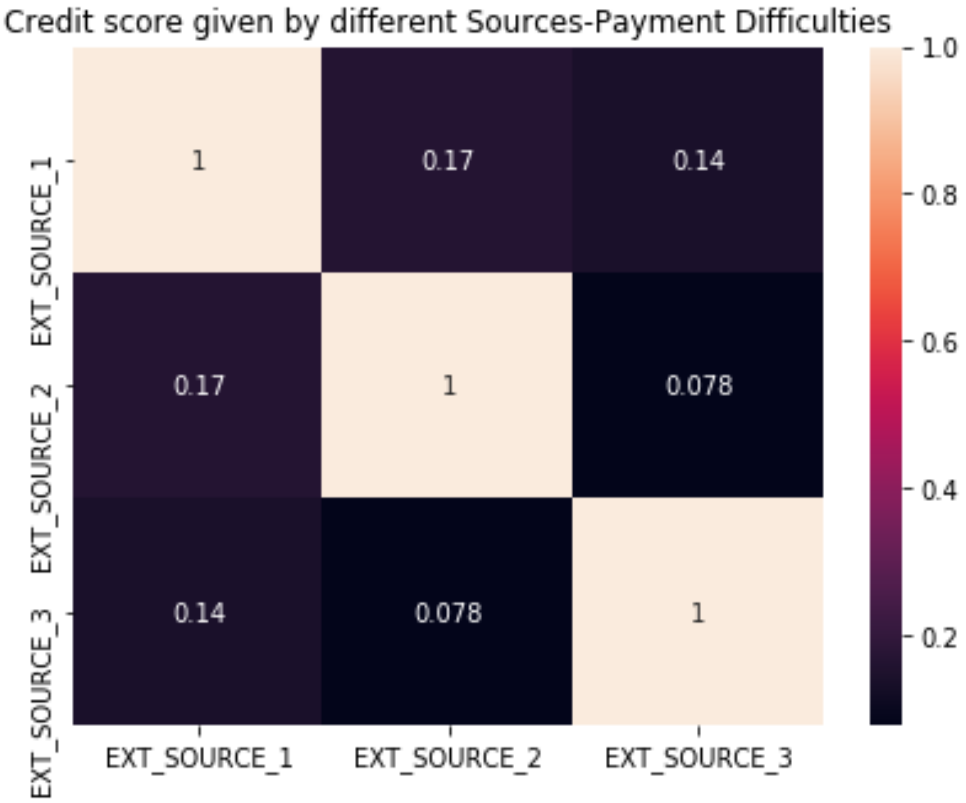


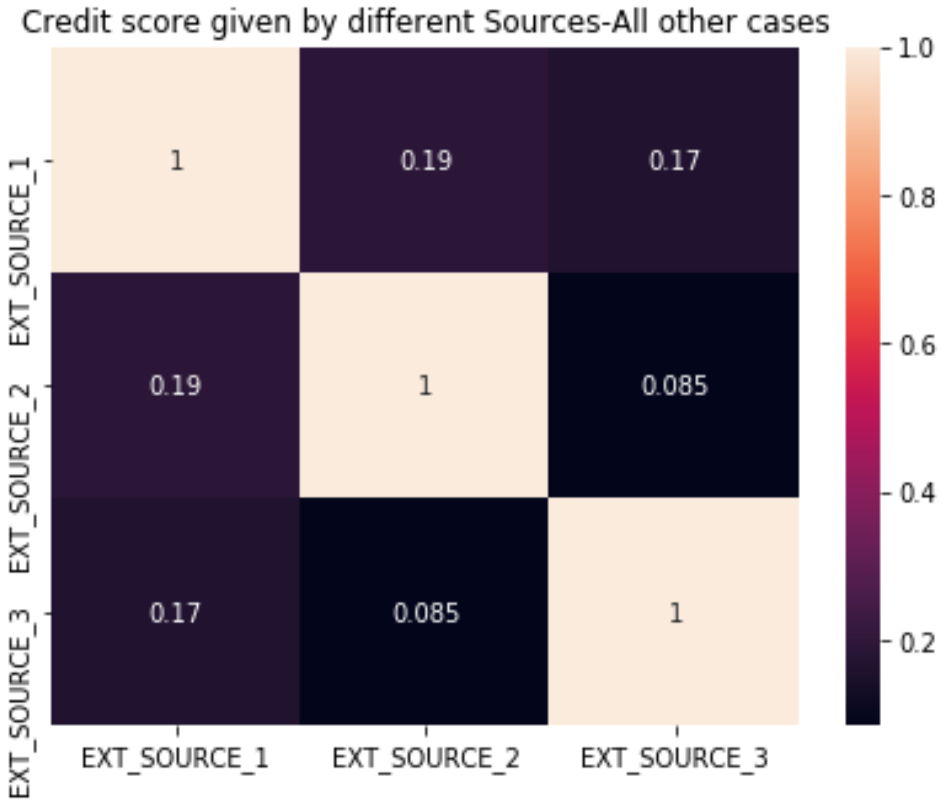
TARGET CUSTOMERS – AMOUNT CREDITED – EDUCATION TYPE



➢ ***From the box plot above we see that the average Credit given to the Clients with academic degree is highest for both who pay late and pay on time , but there is missing data on the total income of such clients. Hence we need to investigate the authenticity of the group, as maximum credit is given to them.***

Correlation between the EXTERNAL SOURCES

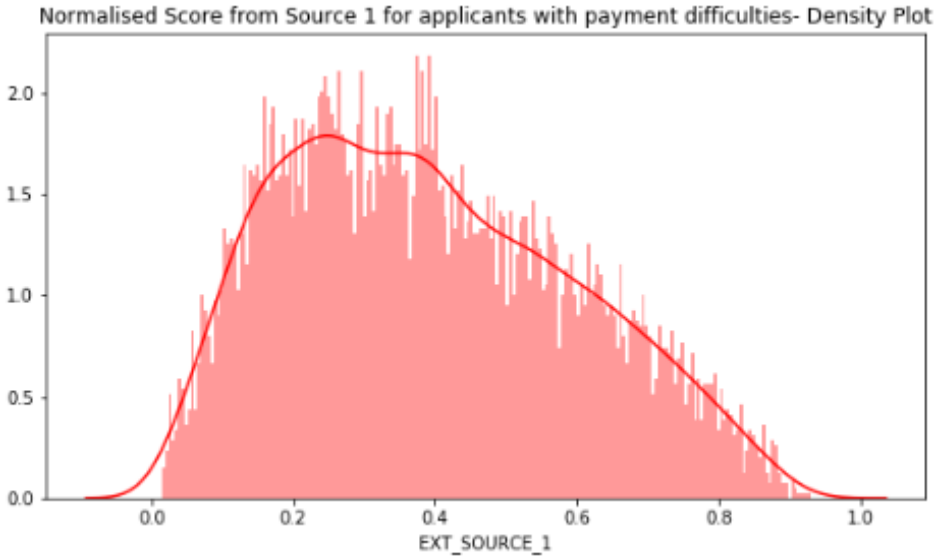


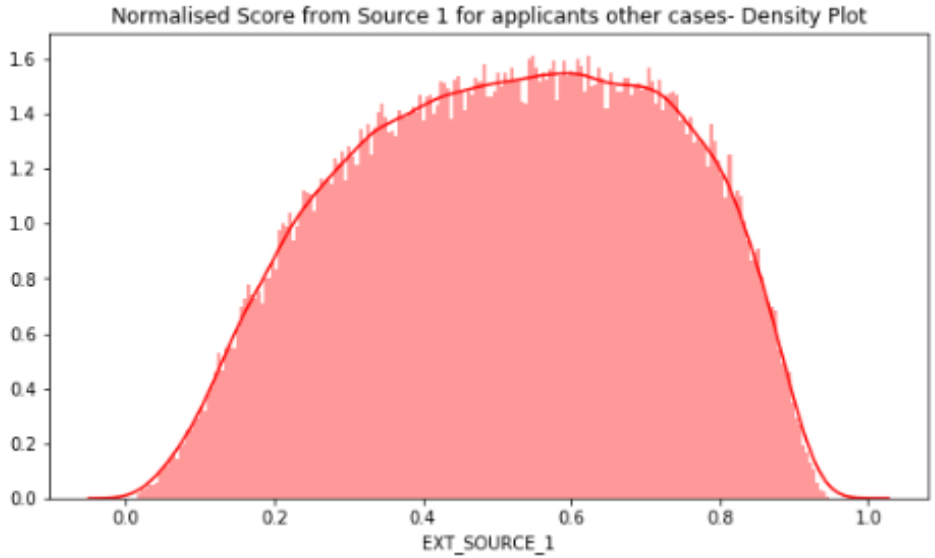


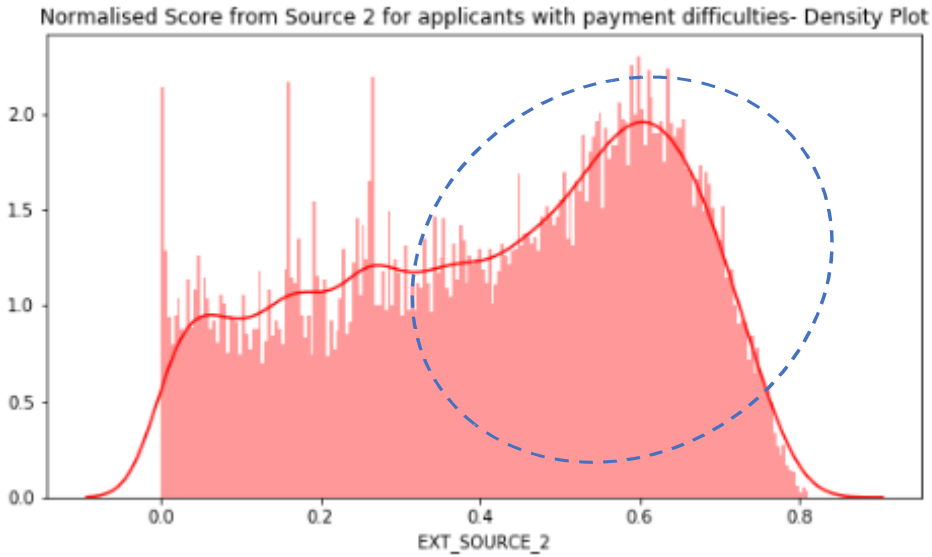
➢ In both the cases i.e for applicants with payment difficulties as well as other cases, we can find very small correlation between all the 3 sources for normalised score out of which External Source 2 and External Source 3 are the least correlated. Hence while finding out for defaulters, only checking on the normalised score provided by these sources won't be efficient, since the scores provided would not be similar

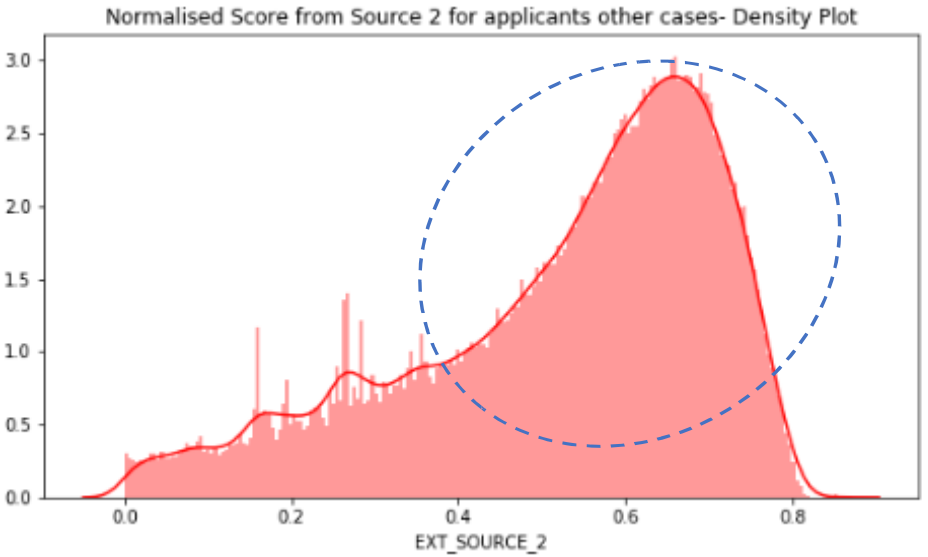
.

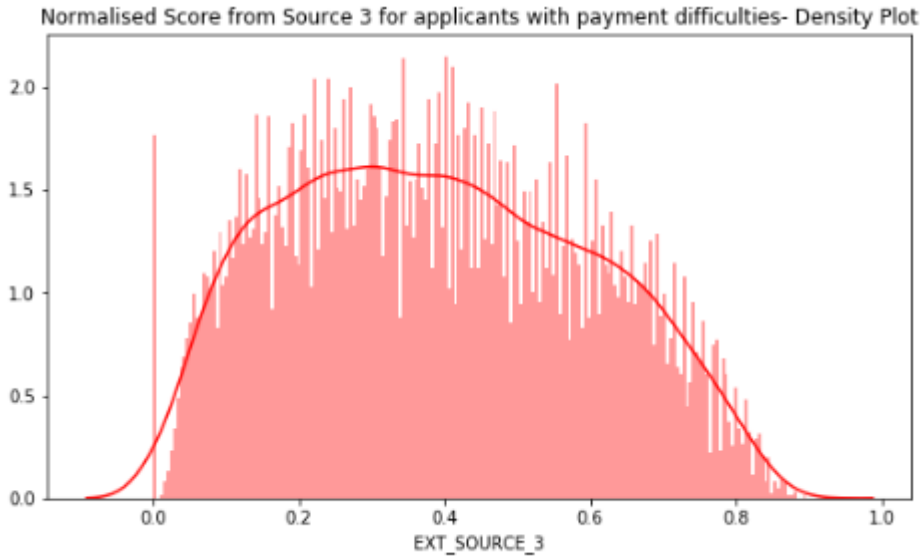
Now Let the DENSITY PLOTS Talk!











INFERENCE:

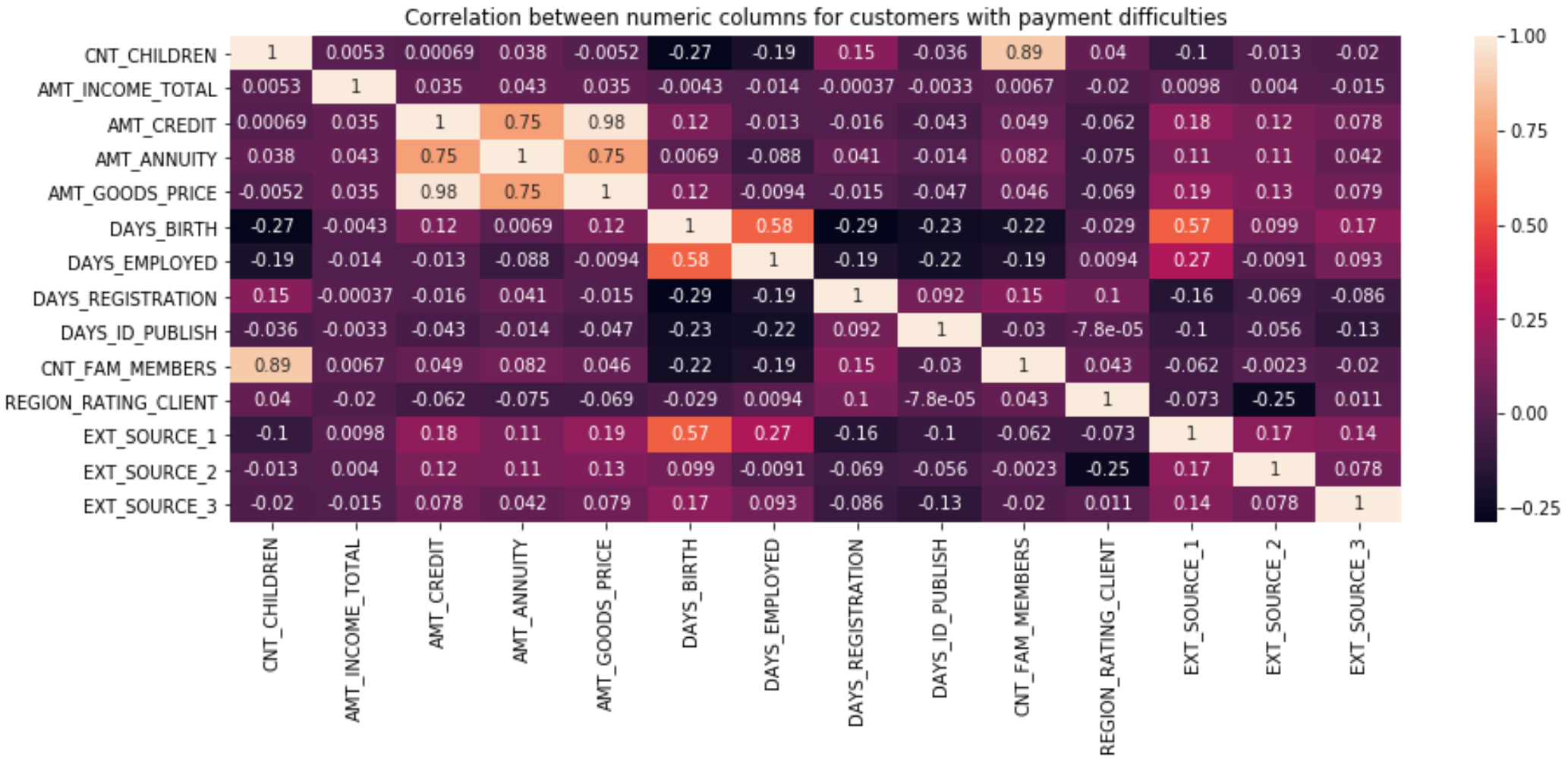
❑From the above Density Plots, we can find for External Source1 the density plot looks kind of inversely proportional for the normalised score for applicants with payment difficulties and the applicants other cases and looks valid since we Expect the Normalised score to be less for applicants with payment difficulties.

❑Whereas, for External Source2, we can find a high density for high normalised scores even for applicants with payment difficulties as well as in other cases. Since the score would be generated based on the previous loan instances of the applicants, this score might show high, as there won't be any previous loan application by the applicant. This can be one case.

❑But in other ways, we can also find from the External Source 1 and 3,where the normalised score is less for applicants with payment difficulty which contradicts the density plot of External Source 2 considering the same data points used. Hence it would be better not to just depend only on the normalised scores for judging the applicant to be defaulter.

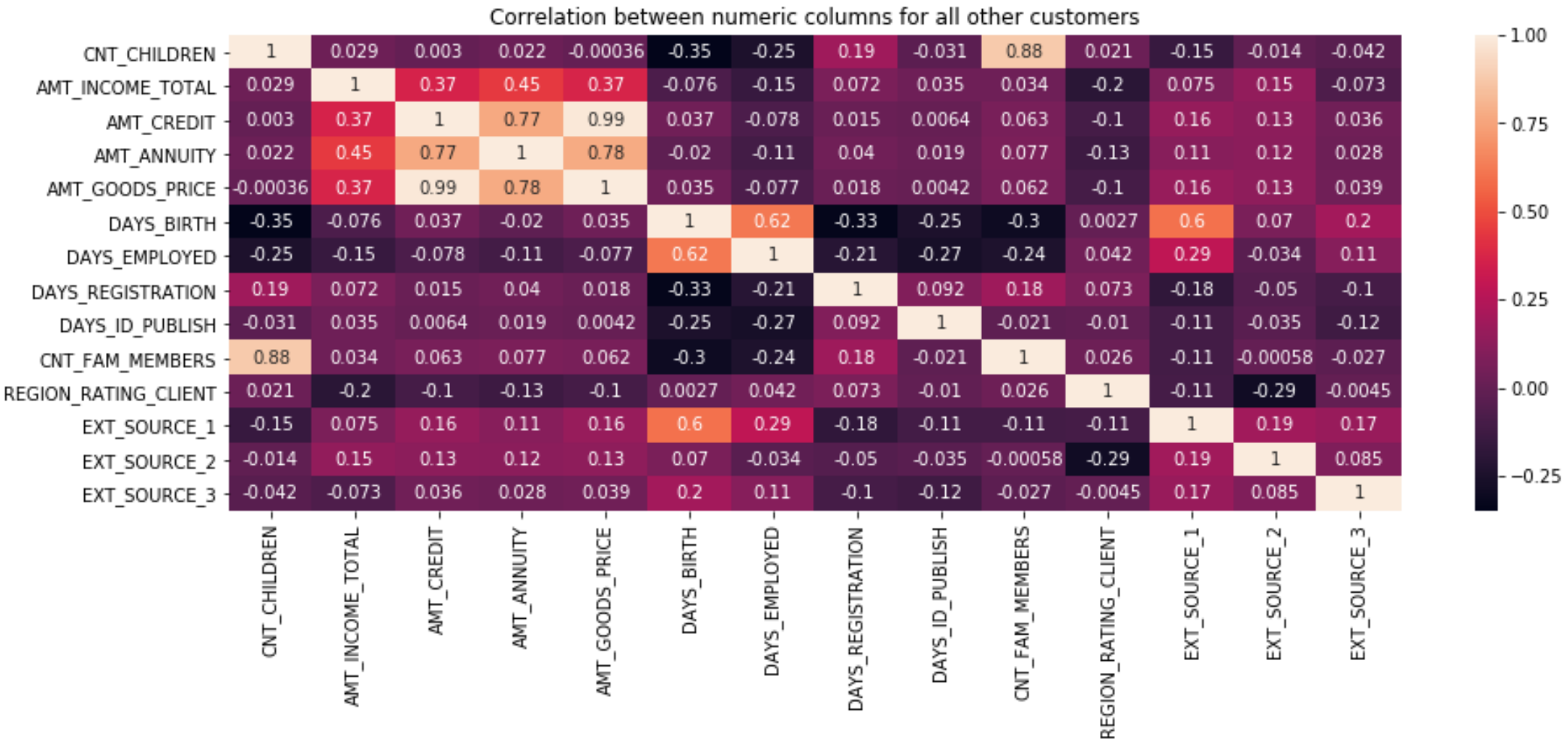
Hence individual density plots gave us an idea why there is less correlation between the different Normalised scores as seen before.

Numerical Columns Correlation – Payment Difficulties



➢ **There is a high correlation between days of Birth and Normalised score from external source 1.** This means the more external source 1 is more inclined/biased towards aged people in terms of providing high credit score.

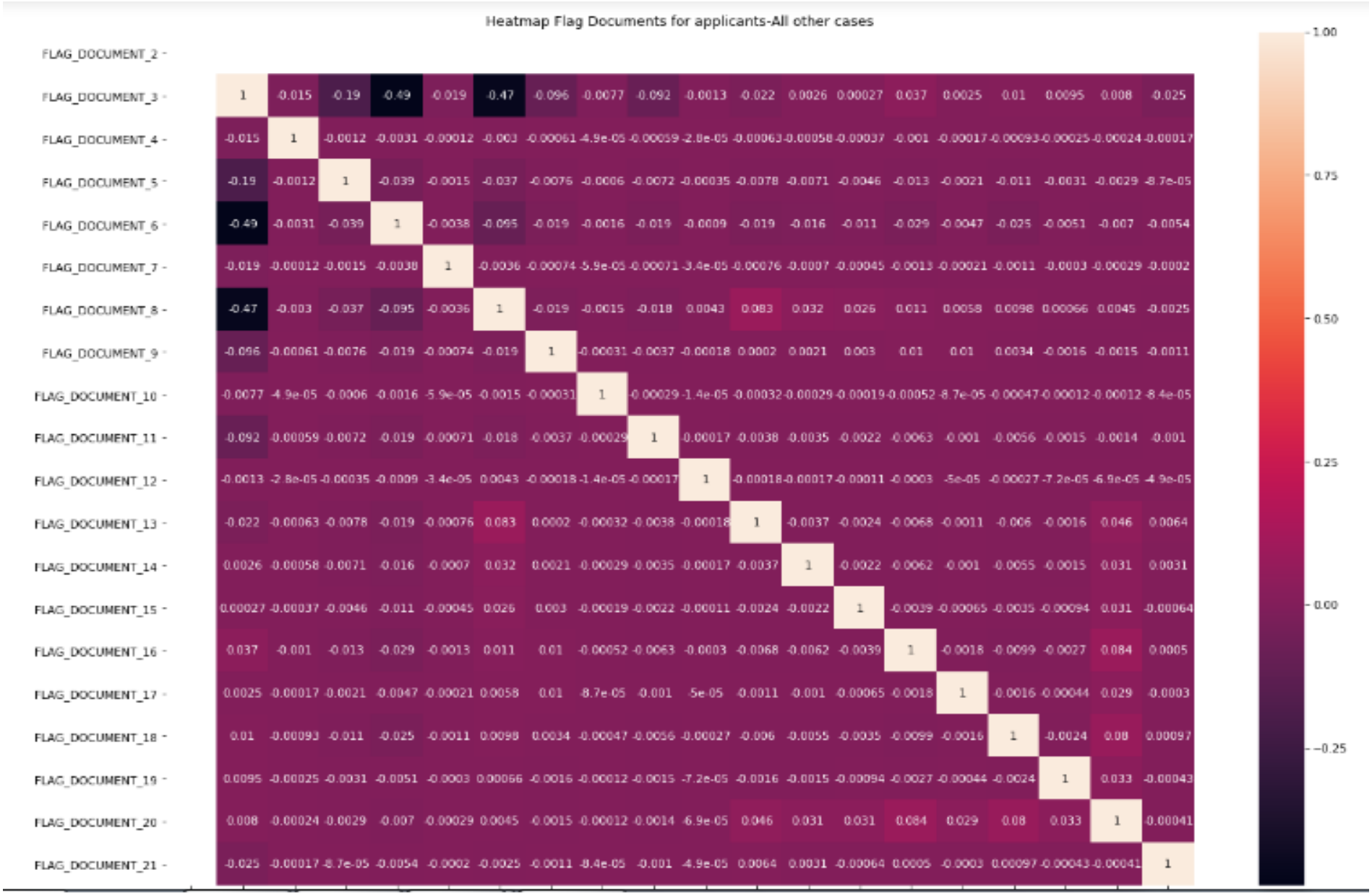
Numerical Columns Correlation – Other Columns



Conclusion

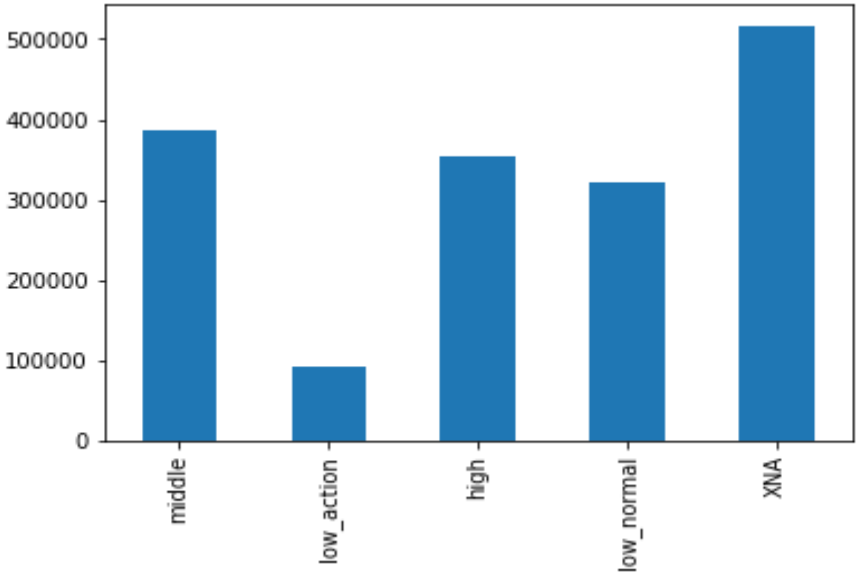
However as correlation does not lead to causation*, if any aged people have high credit score, it is not likely that it is due to external source 1.*

Any Relation between the FLAG DOCUMENTS?? – All Other Cases



WHAT PREVIOUS APPLICATION DATA SHOWS US?

INTEREST RATE GROUPS



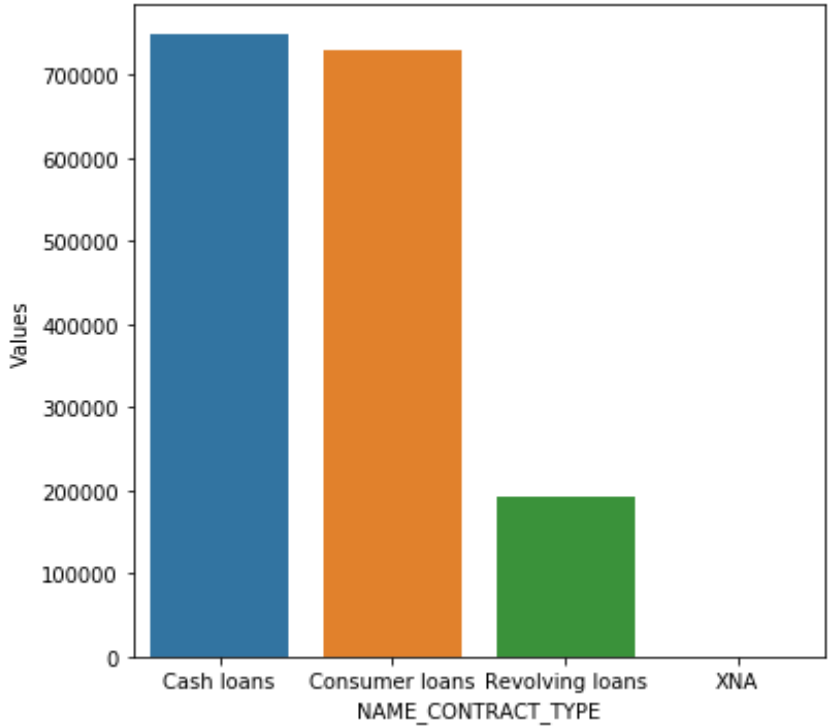
➢ Ignoring the XNA category values, we can find Interest rate groups are usually high or in middle category. Low-action Interest rate group is less.

Loan Types for the Previous Applications

➢Most of the applicants previous Loans are Cash Loans or Consumer Loans.

➢Revolving Loans are the least provided Loan Type.

.



ANNUITY AMOUNT – APPLICATION AMOUNT – CREDIT AMOUNT – GOODS AMOUNT

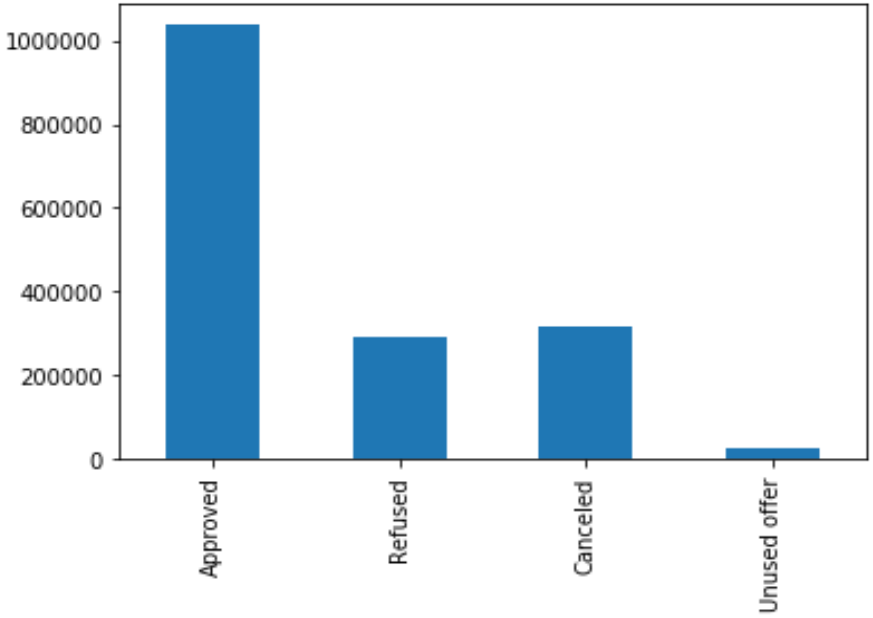
➢The Goods Price Amount, Amount Annuity, Application Amount and the Credit Amount are all highly correlated with each there as expected.

.

REFUSED APPLICATIONS

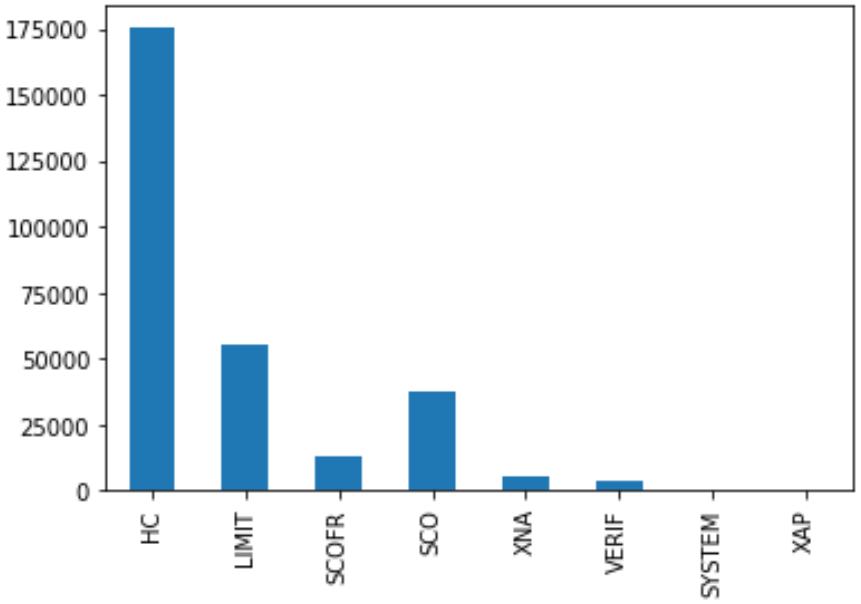
```
prev_app_refused.shape[0]
```

```
290678
```



➢ 2.9 Lakh of the client applications where refused by the company for Loan out of the 16.7 Lakh applications received.

REJECT REASON FOR APPLICANTS?



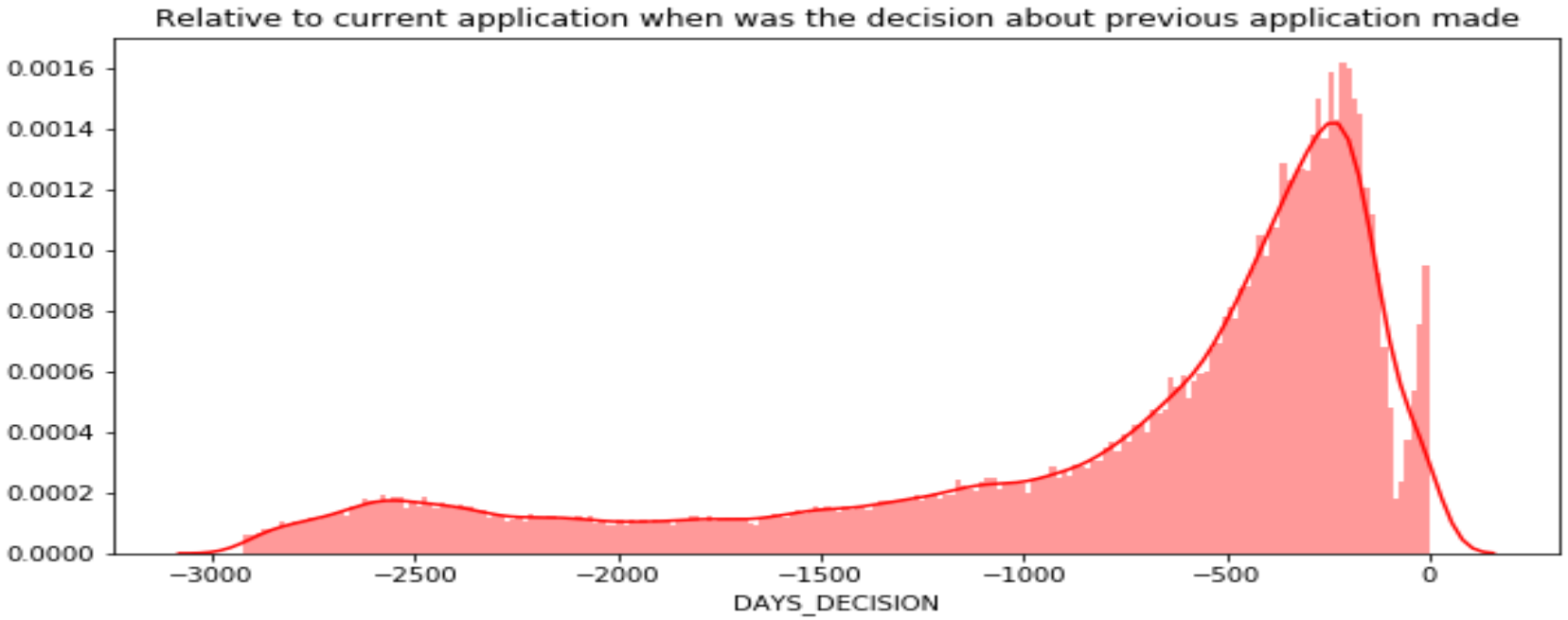
➢**From the Rejected applicants, the Reject Reason for most of the applicants is HC and LIMIT.**

.

Outlier Detection

➢ From the Box plot we can find many outliers are present with respect to the Application amount requested by the client, with one application standing out **with almost 70Lakh as requested amount**. Also it is clear that a major part of the Requested amounts are actually maybe less than 5Lakhs.

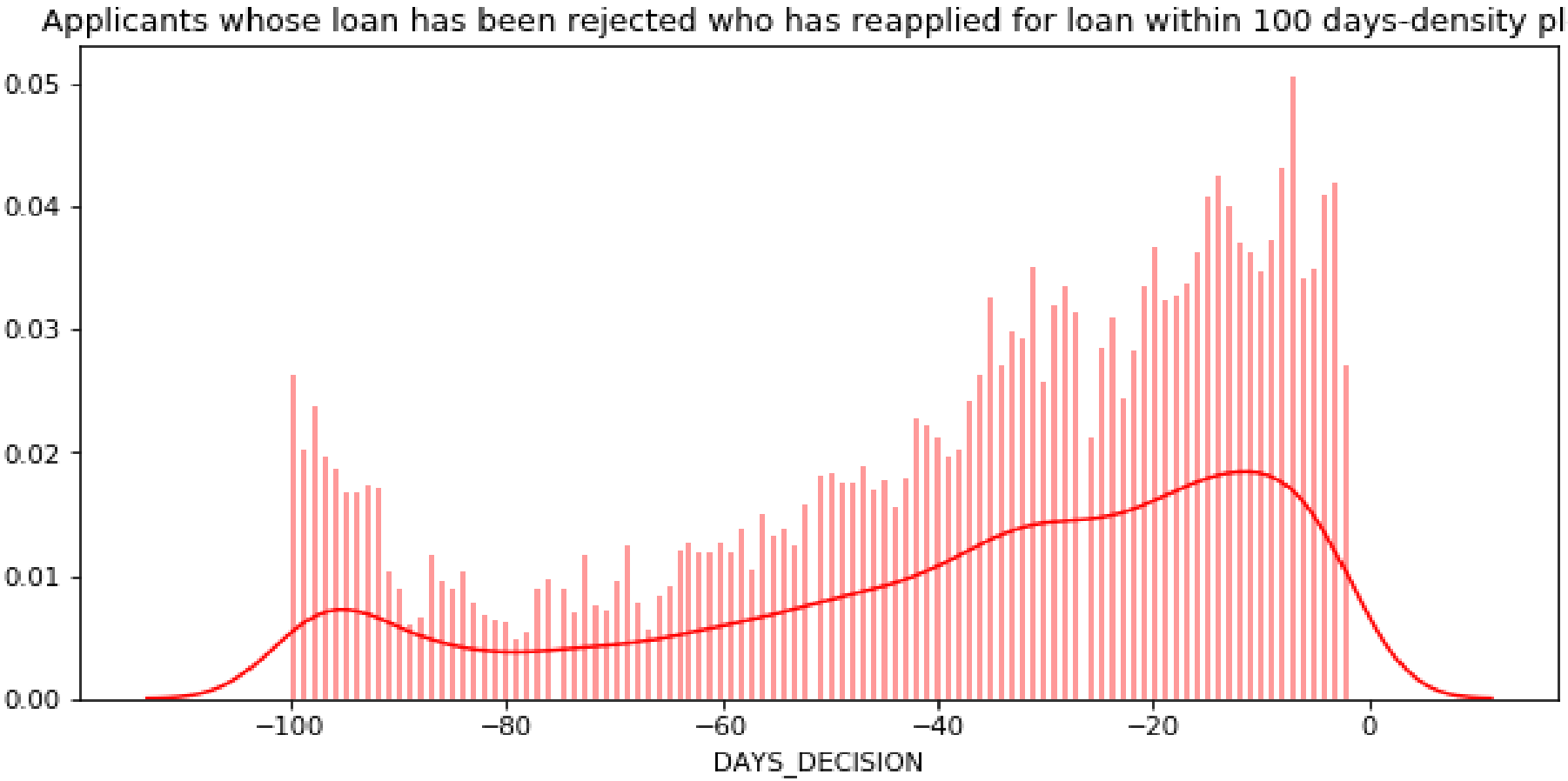
Current Application and the Decision about the Previous Application –for REFUSED APPLICATIONS



➢ From the above Plot we can find that most of the current applicants whose previous application was rejected as reapplied for a Loan within a minimum of 500 days and many number of people have reapplied the day their previous application was rejected itself.

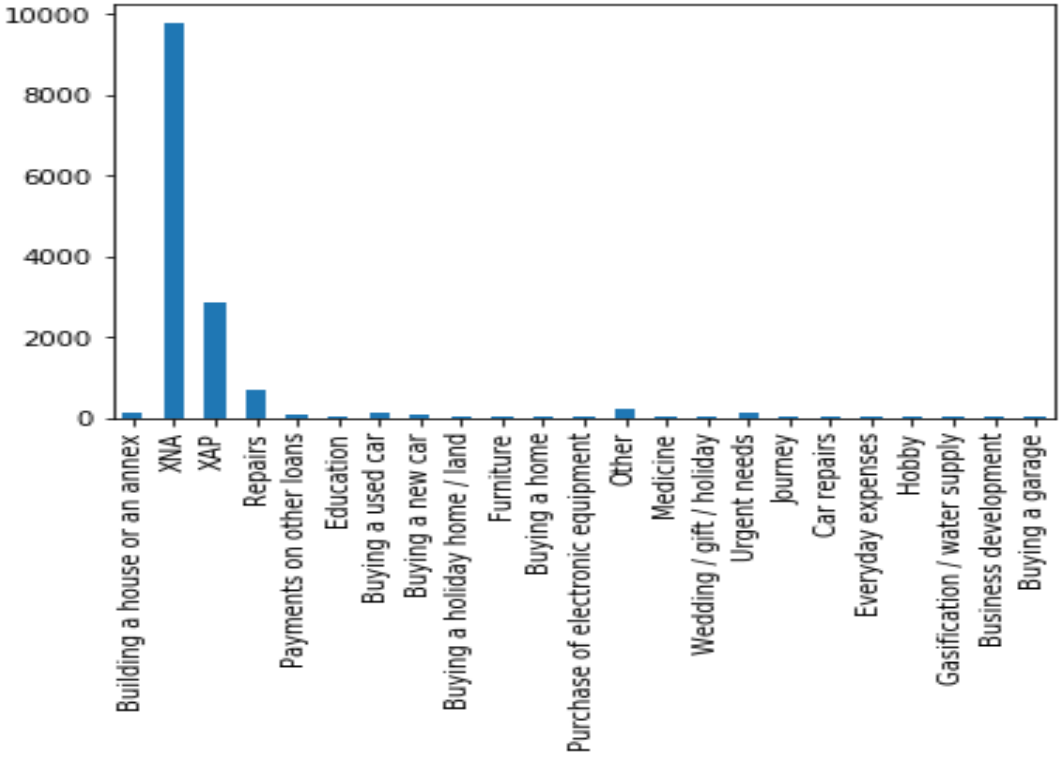
➢**Days Decision can be taken as an important category while lending out loans for the applicants.**

Current Application and the Decision about the Previous Application –for REFUSED APPLICATIONS who REAPPLIED IN 100 DAYS



➢ It is most likely that if an applicant applies loan within **20 days** of previous loan rejection, it is highly likely that this application will also get rejected as per the analysis above

Current Application and the Decision about the Previous Application –for REFUSED APPLICATIONS who REAPPLIED IN 100 DAYS



- **Almost 68% of the people as mentioned above as not stated purpose for the Loan. It would be a good practice for the bank to take note of such cases and be cautious in lending loans to such applicants.**

```
(refused_days_less_thanyear.loc[refused_days_less_thanyear['NAME_CASH_LOAN_PURPOSE']== 'XNA']['NAME_CASH_LOAN_PURPOSE']).count()
```

```
9763
```

```
refused_days_less_thanyear['NAME_CASH_LOAN_PURPOSE'].count()
```

```
14255
```

```
round((9763/14255)*100,2)
```

```
68.49
```

INFERENCES

- Thus from the entire Analysis above we see that the data is skewed or more biased towards the one without payment difficulties than others. In terms of Types of Customers belonging to Working categories, more samples of Labourers are taken into account during survey.

- Also, it is seen that, most of the people in the group are from having average income, hence the customers having payment difficulties might not be facing the issue due to income, it can be related to mode of payment or other expenses.

- Also, for about 68% of the customers bank does not have a proper reason for the loan given. They should also engage more customers in revolving loans, which kind of improves customer retention.

- Thus bank should focus on the above areas to reduce number of defaulters and retain loyal.

THANK YOU