CREDIT EDA CASE STUDY

BY

RAHUL KUMAR GUPTA

Introduction

- This assignment aims to give an idea of applying EDA in a real business scenario. In this assignment, a basic understanding of risk analytics in banking and financial services is used and understand how data is used to minimise the risk of losing money while lending to customers.

Business Understanding

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected. When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

- If the applicant is not likely to repay the loan, i.e., he/she is likely to default, then approving the loan may lead to a financial loss for the company

Business Understanding - 2

- The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- **The client with payment difficulties**: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- **All other cases**: All other cases when the payment is paid on time.

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

  1. **Approved**: The Company has approved loan Application
  2. **Cancelled**: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
  3. **Refused**: The company had rejected the loan (because the client does not meet their requirements etc.).
  4. **Unused offer**: Loan has been cancelled by the client but on different stages of the process.

- In this case study, we will use EDA to understand how consumer attributes and loan attributes influence the tendency of default

Business Objectives

- 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. Identification of such applicants using EDA is the aim of this case study.

- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e., the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

- To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

Data Understanding

- This dataset has 3 files as explained below:

- 1. inp0 : Contains the information of the client at the time of application.The data is about whether a client has payment difficulties.

- 2. inp1 : Contains merged information about the client's current and previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.

Analysis of the information of the client at the time of application.

Missing Values over 13%.

OCCUPATION_TYPE	31.345545
EXT_SOURCE_3	19.825307
AMT_REQ_CREDIT_BUREAU_YEAR	13.501631
AMT_REQ_CREDIT_BUREAU_QRT	13.501631
AMT_REQ_CREDIT_BUREAU_MON	13.501631
AMT_REQ_CREDIT_BUREAU_WEEK	13.501631
AMT_REQ_CREDIT_BUREAU_DAY	13.501631
AMT_REQ_CREDIT_BUREAU_HOUR	13.501631

Treatment of Missing Values

- 0.0 71801
- 1.0 63405
- 2.0 50192
- 3.0 33628
- 4.0 20714
- 5.0 12052
- 6.0 6967
- 7.0 3869
- 8.0 2127
- 9.0 1096
- 11.0 31
- 12.0 30
- 10.0 22
- 13.0 19
- 14.0 10
- 17.0 7
- 15.0 6
- 19.0 4
- 18.0 4
- 16.0 3
- 25.0 1
- 23.0 1
- 22.0 1
- 21.0 1
- 20.0 1
- Name: AMT_REQ_CREDIT_BUREAU_YEAR, dtype: int64

0.0 215417
1.0 33862
2.0 14412
3.0 1717
4.0 476
5.0 64
6.0 28
8.0 7
7.0 7
261.0 1
19.0 1
Name: AMT_REQ_CREDIT_BUREAU_QRT, dtype: int64

0.0 222233
1.0 33147
2.0 5386
3.0 1991
4.0 1076
5.0 602
6.0 343
7.0 298
9.0 206
8.0 185
10.0 132
11.0 119
12.0 77
13.0 72
14.0 40
15.0 35
16.0 23
17.0 14
18.0 6
19.0 3
24.0 1
23.0 1
27.0 1
22.0 1
Name: AMT_REQ_CREDIT_BUREAU_MON, dtype: int64

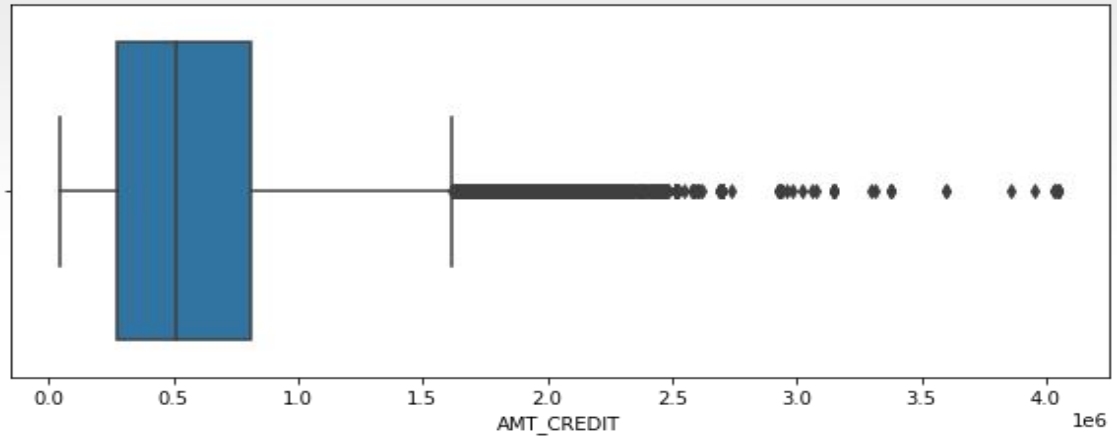
0.0 257456
1.0 8208
2.0 199
3.0 58
4.0 34
6.0 20
5.0 10
8.0 5
7.0 2
Name: AMT_REQ_CREDIT_BUREAU_WEEK, dtype: int64

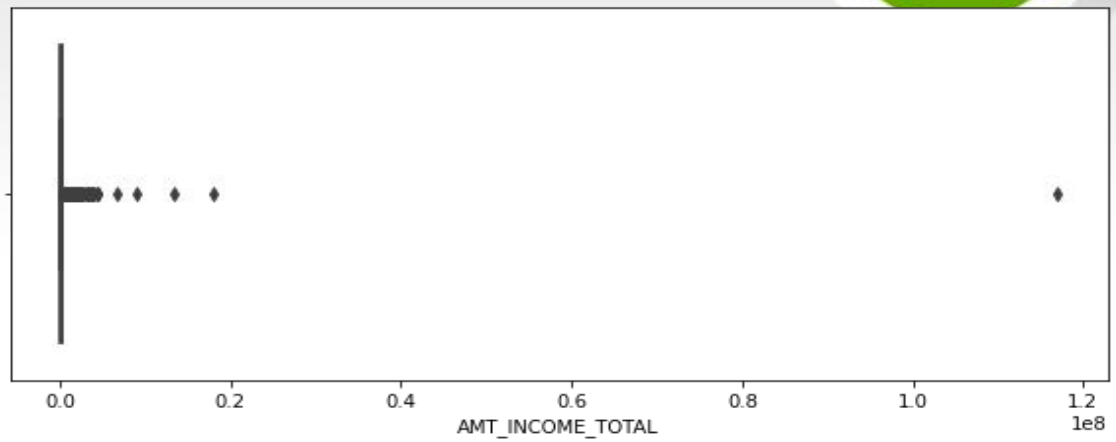
0.0 264503
1.0 1292
2.0 106
3.0 45
4.0 26
5.0 9
6.0 8
9.0 2
8.0 1
Name: AMT_REQ_CREDIT_BUREAU_DAY, dtype: int64

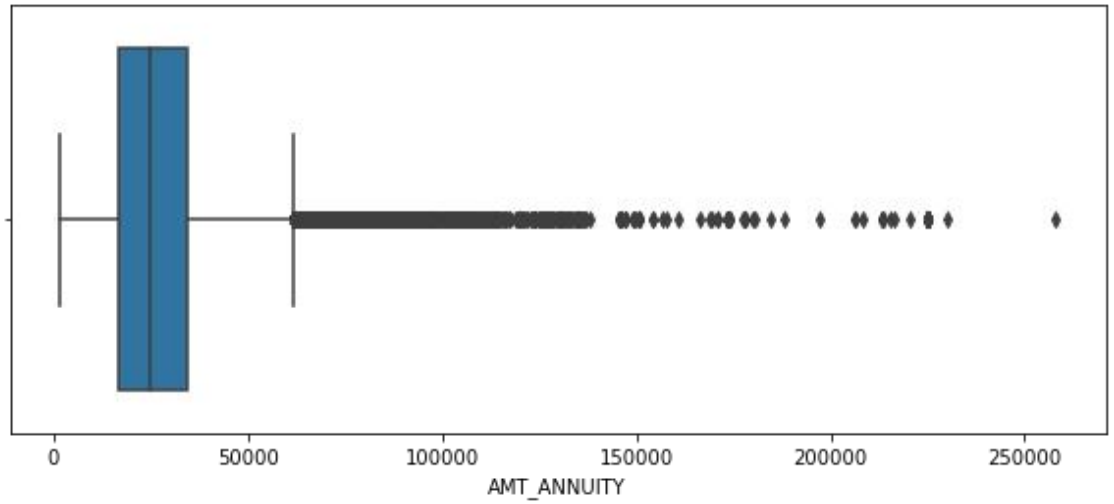
0.0 264366
1.0 1560
2.0 56
3.0 9
4.0 1
Name: AMT_REQ_CREDIT_BUREAU_HOUR, dtype: int64

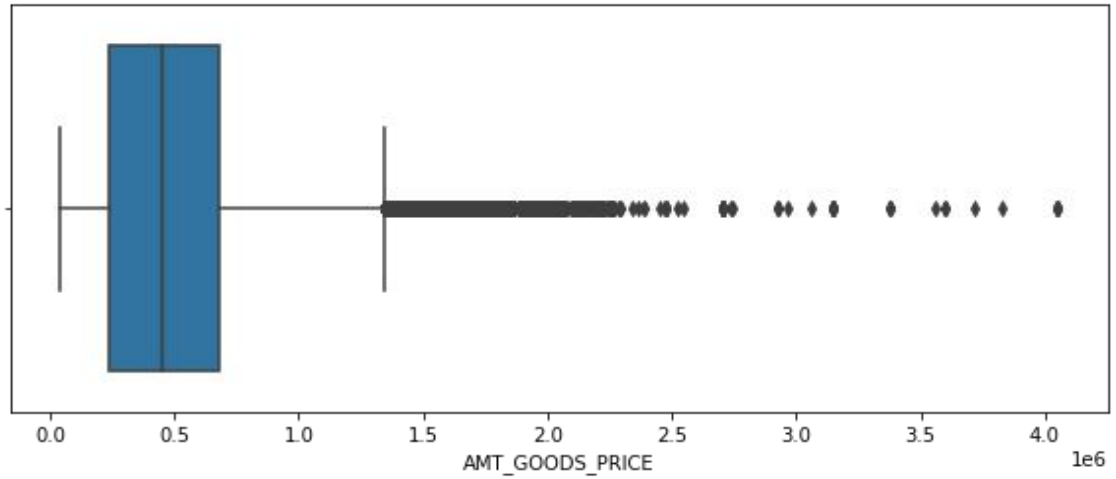
* As We can see that AMT_REQ_CREDIT_BUREAU columns are in Days,Month,year and so can be imputed with mode or 0.
And the wrong value (261) can too be imputed with mode or 0.

OUTLIERS



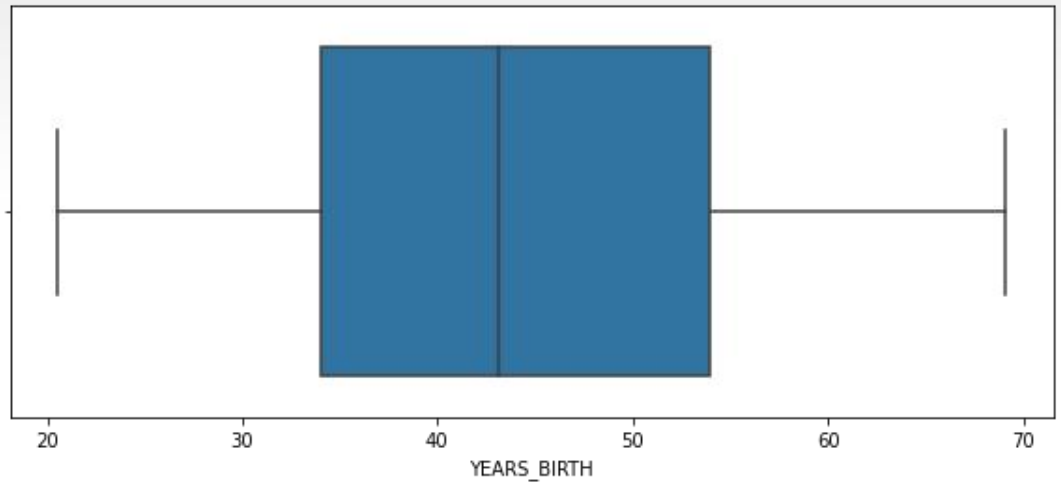


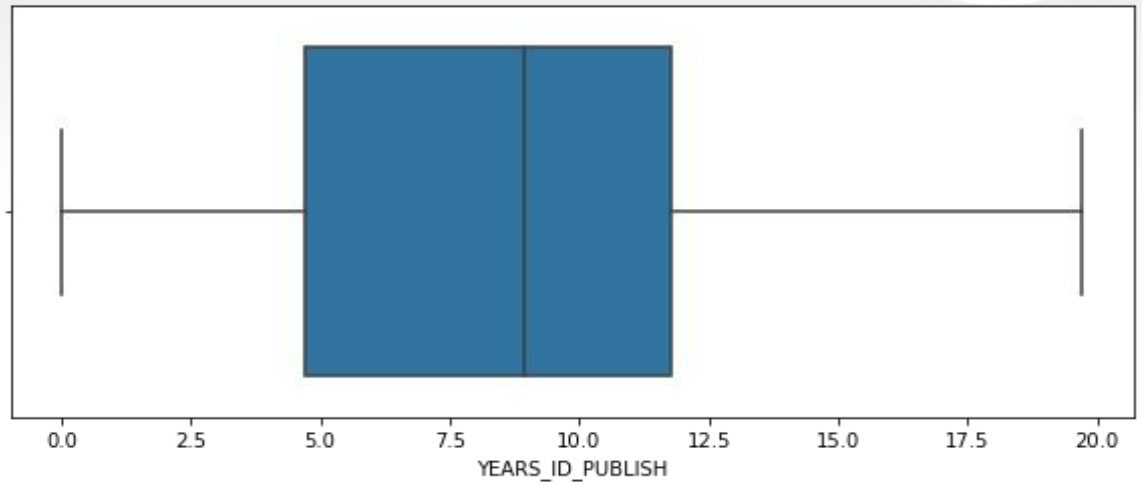


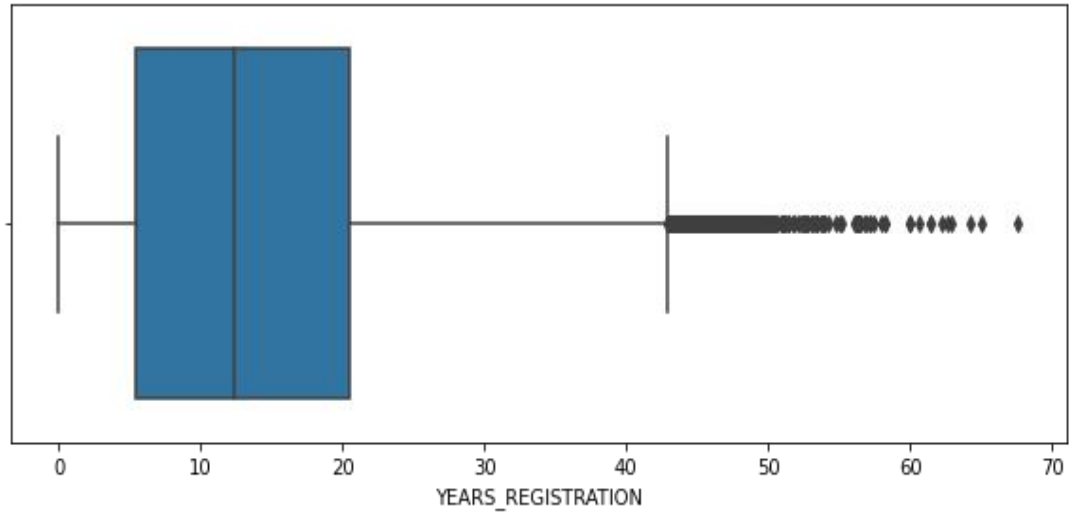


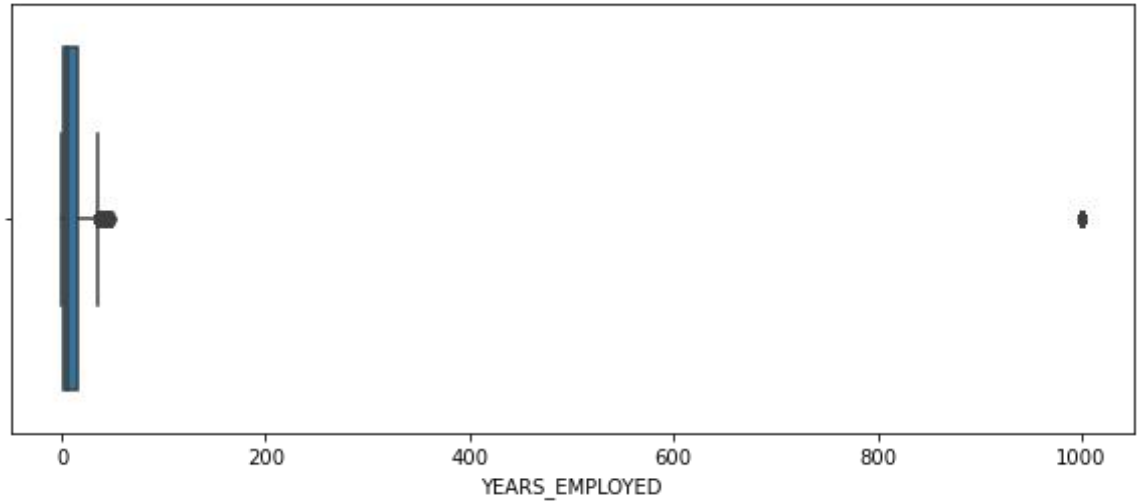
We see that AMT_INCOME_TOTAL has an outlier which is too far from the group near 1.2.
We can see that Credit also has outliers around 3 to 4.2
Annuity also has outliers around 200000 to 260000.
Goods price also show outliers at 2.7 to 4.1

Outliers



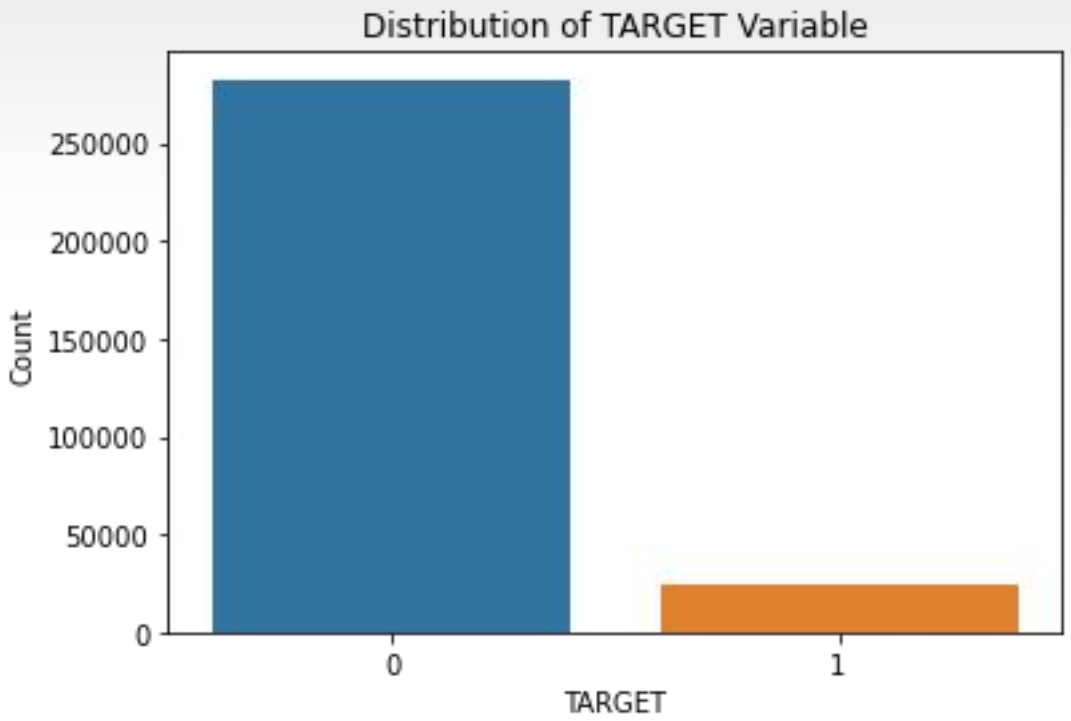


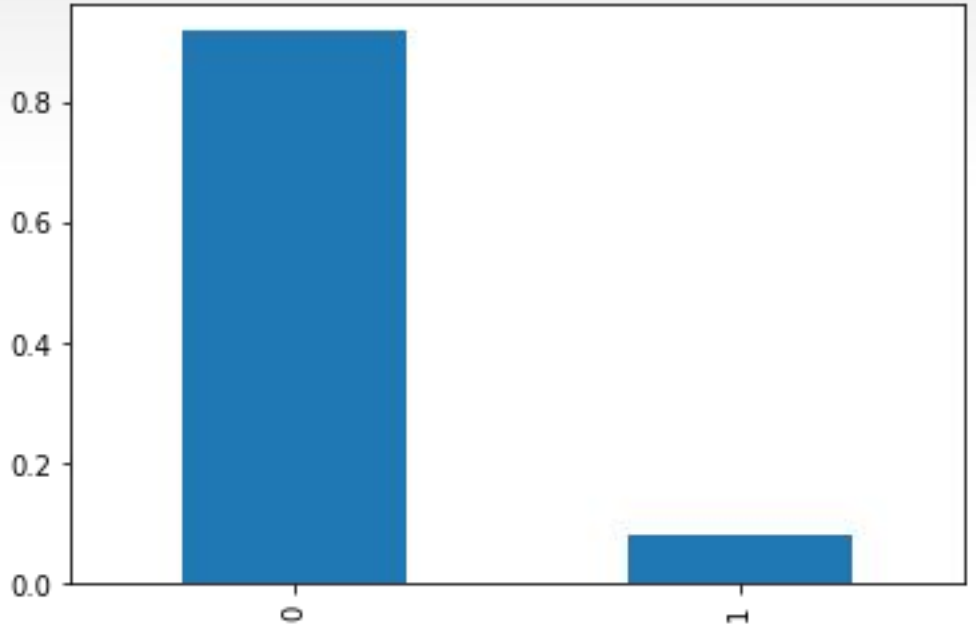




Years Employed shows an outlier at 1000 sine it is not possible to work for 1000 years.

Analysis Of TARGET Variable





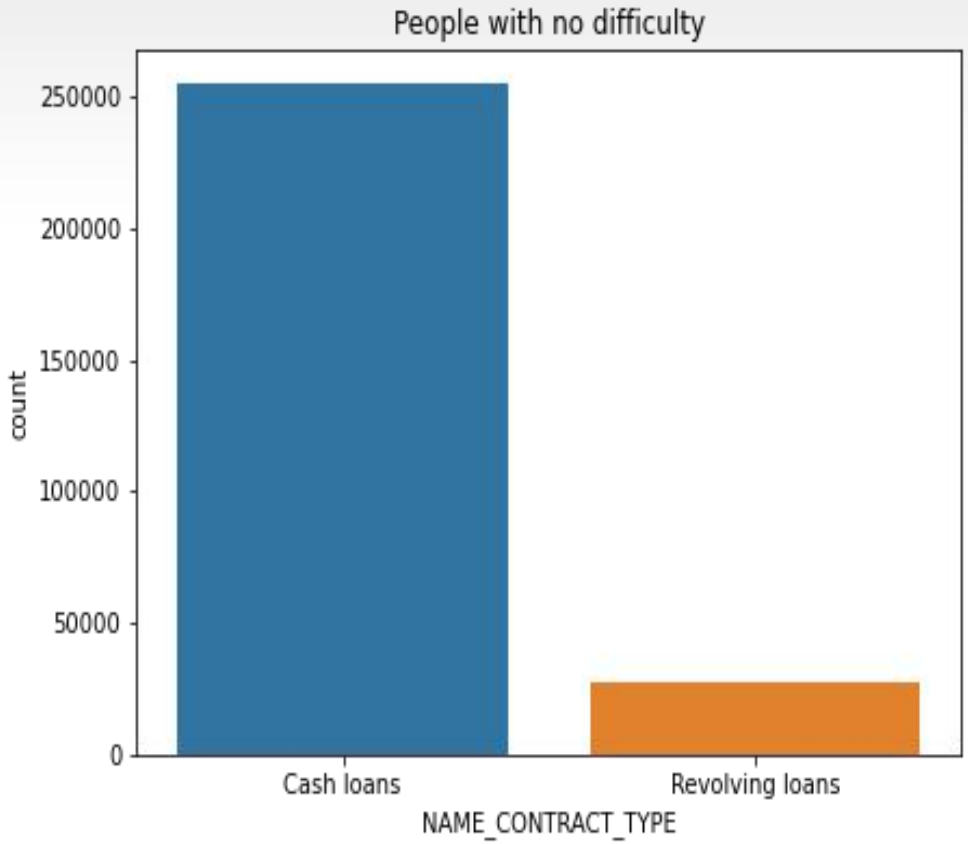
0 282668
1 24823
Name: TARGET, dtype: int64

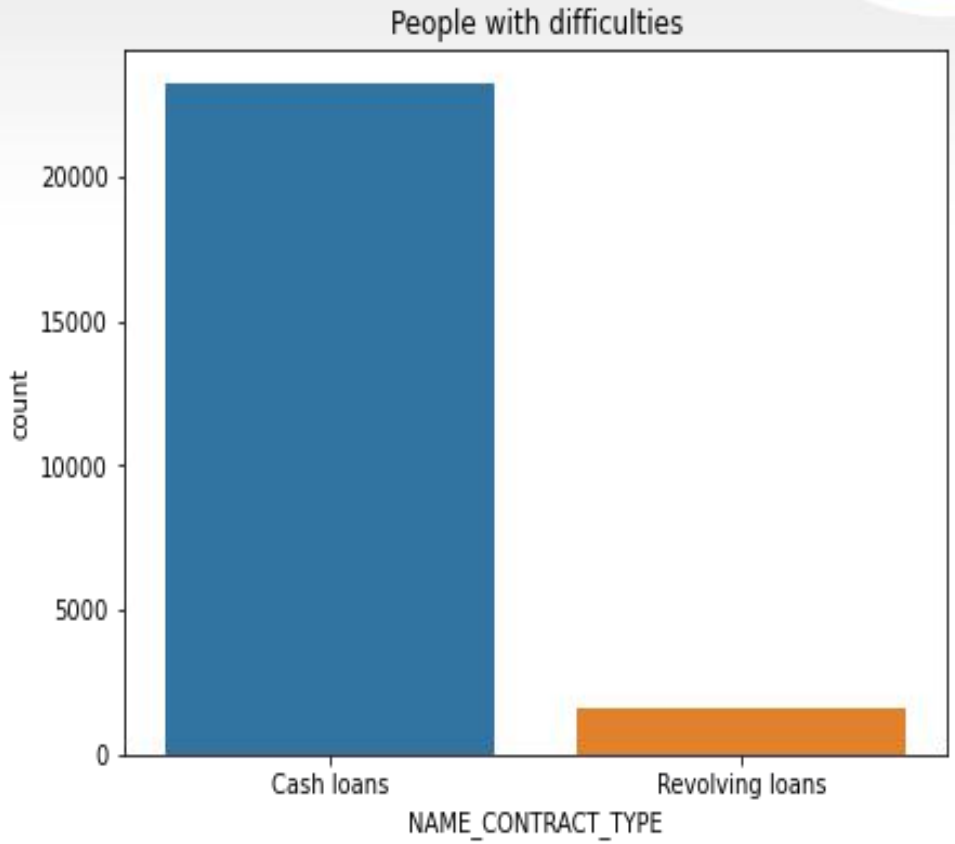
0 91.927243
1 8.072757
Name: TARGET, dtype: float64

There is a huge data imbalance.
The People with no difficulty in Paying are 91%.
The People who default are only 8.07%.

UNIVARIATE ANALYSIS

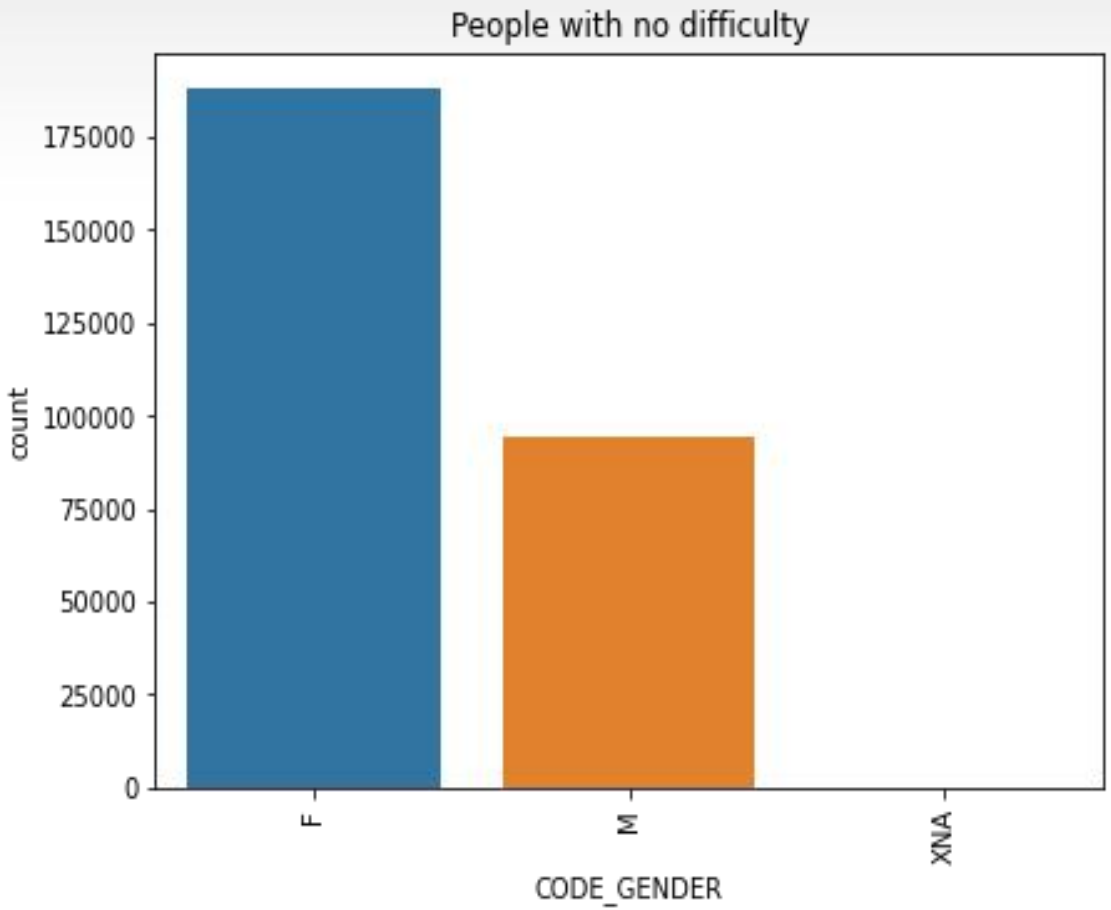
TYPE OF LOAN

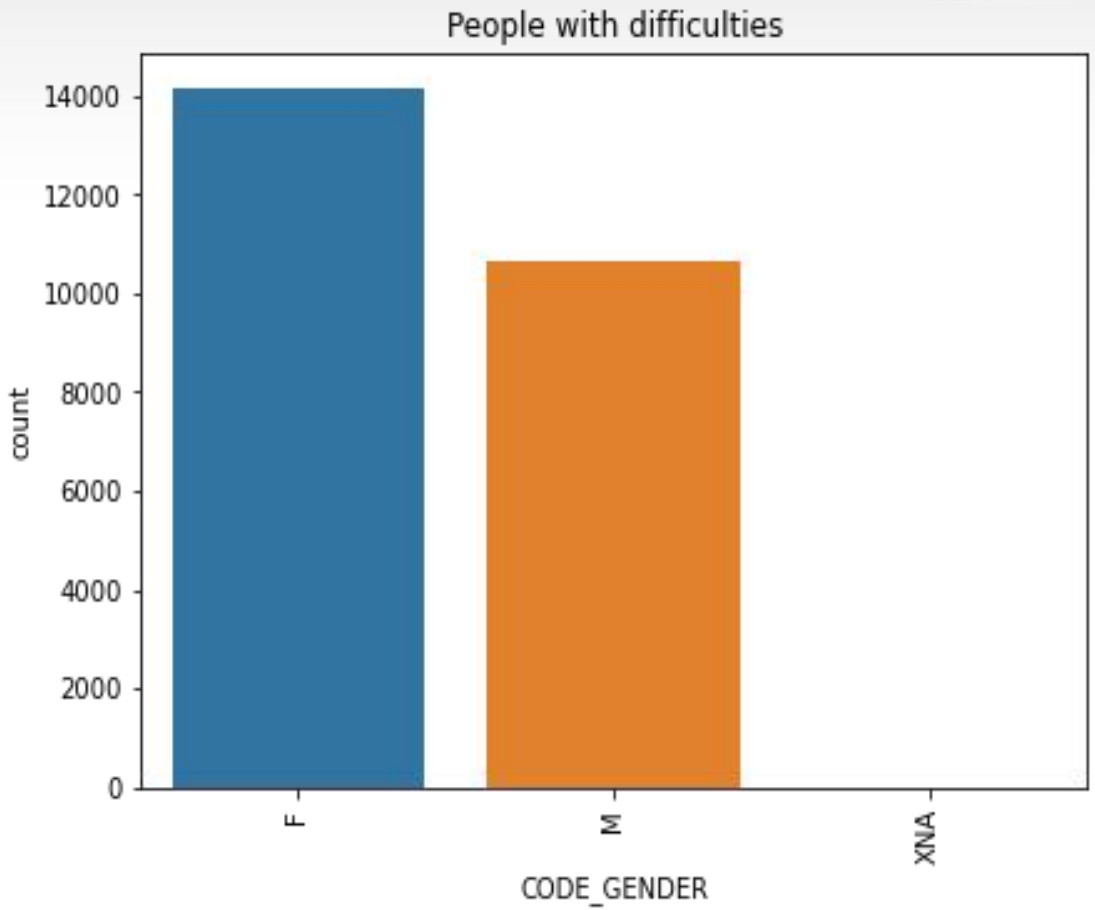




* We can see that most of the loans are Cash loans for both(almost 90%). Revolving loans are 3% more in People with no difficulties.

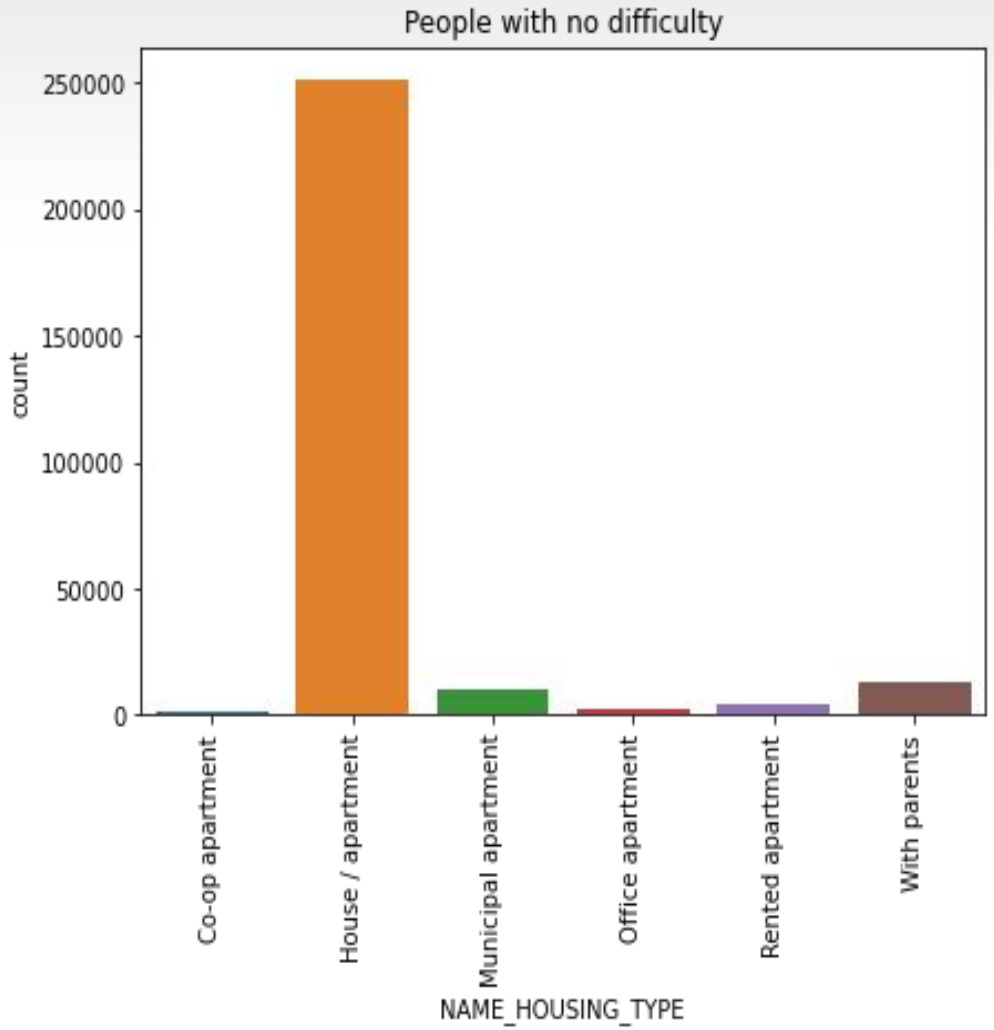
Gender

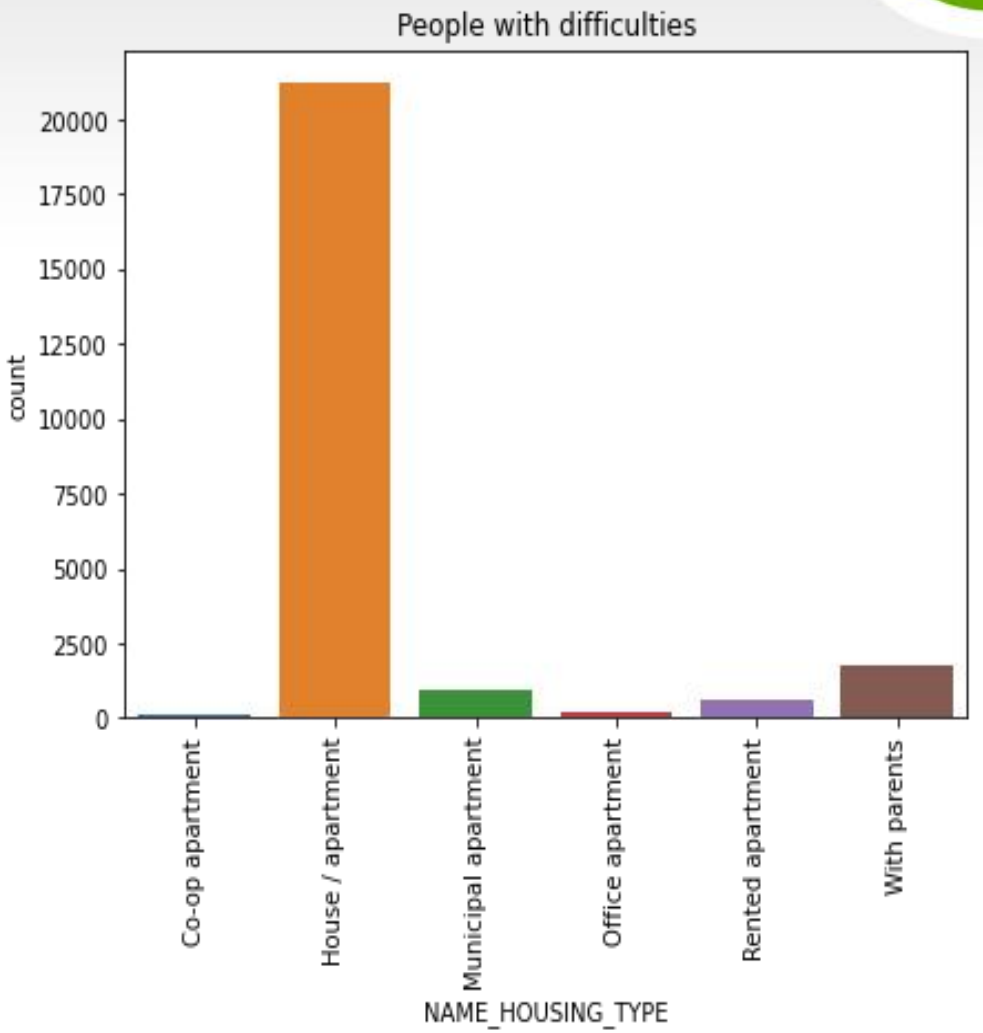




We can see here that Females have taken more loans than males.
But there are more males in difficulty than with no difficulty.

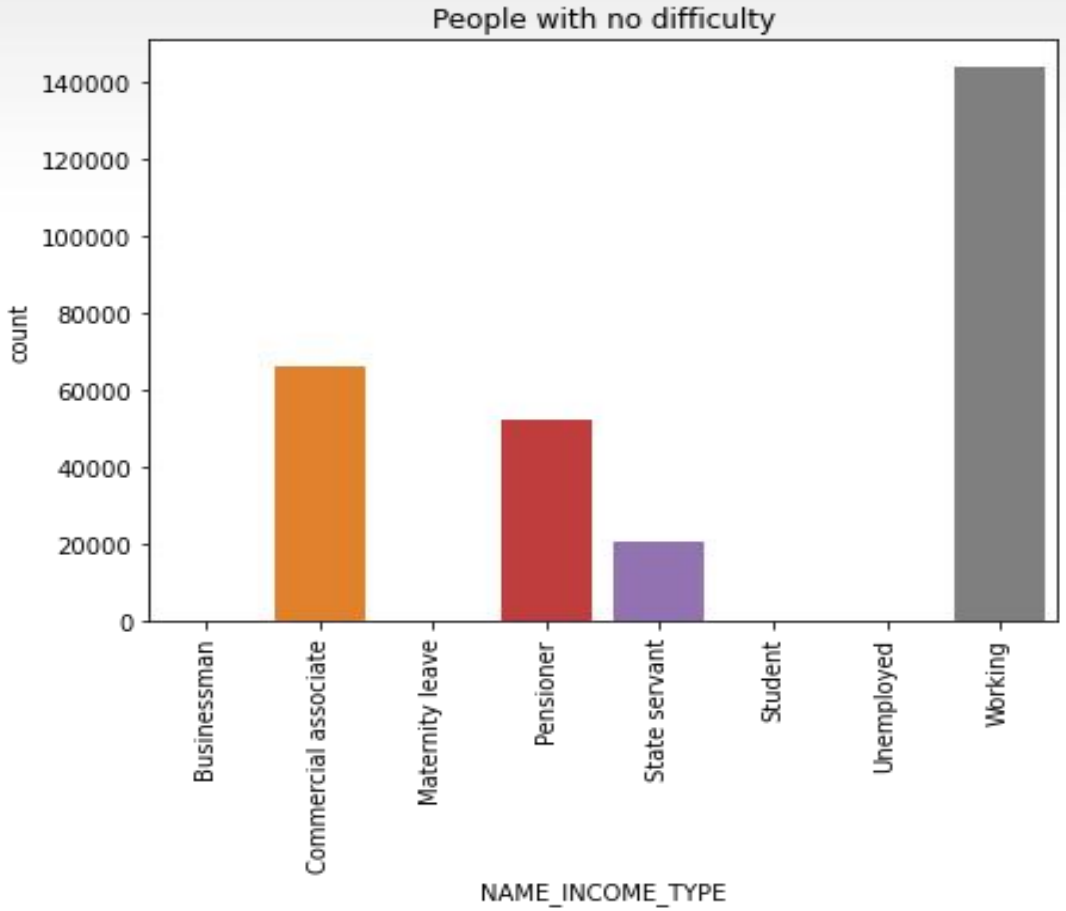
Housing

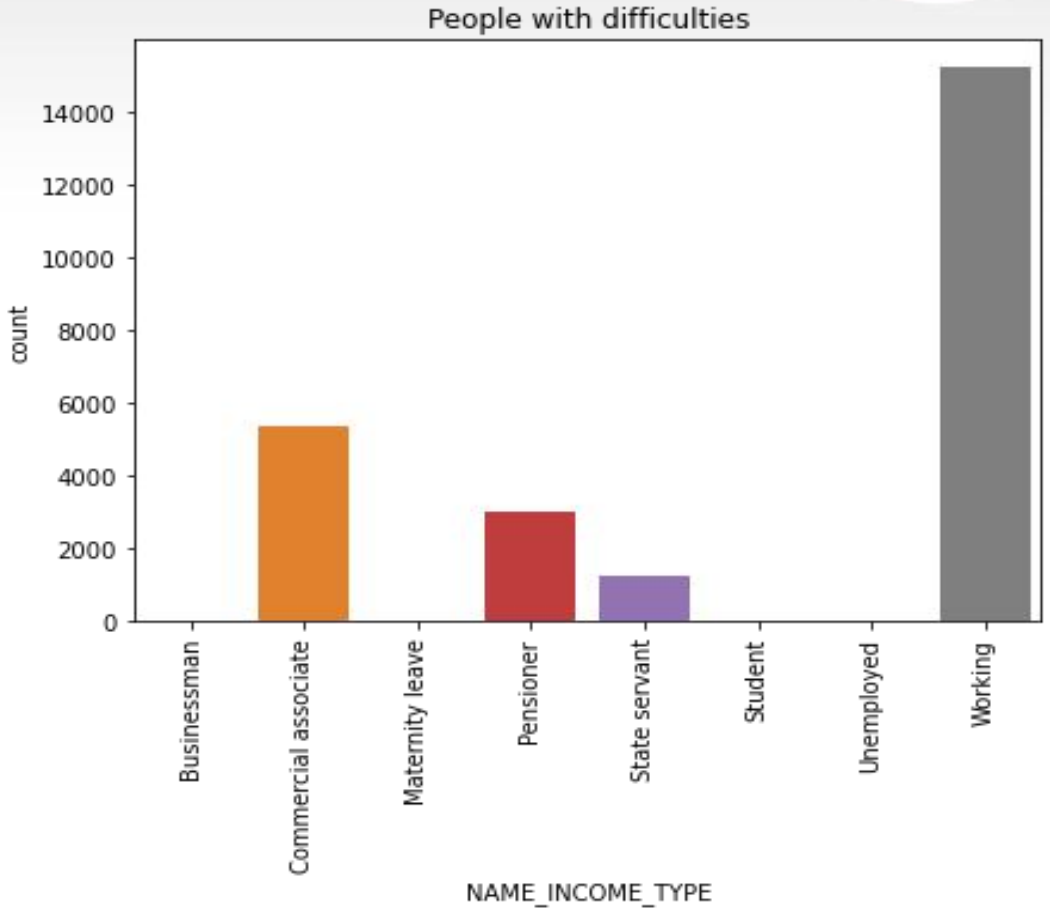




We see that in both, people living in House/apartment are more.

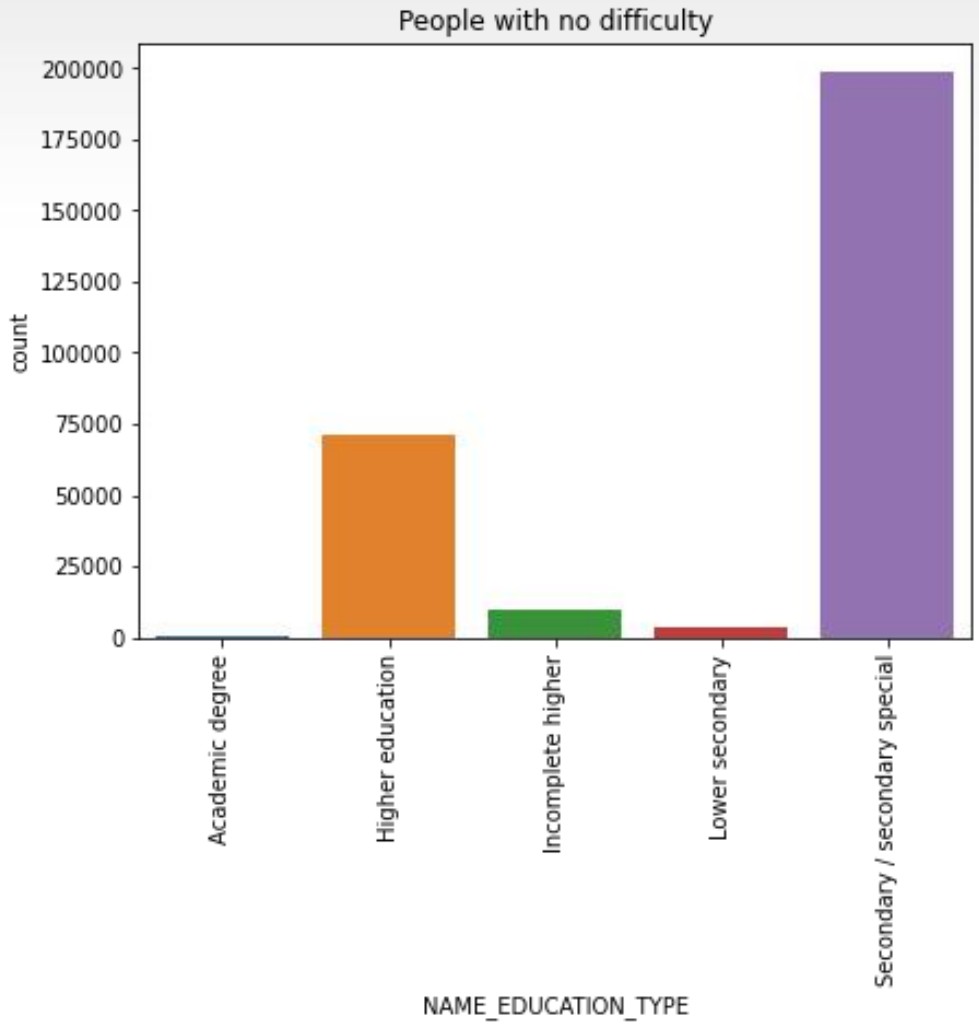
Income Type

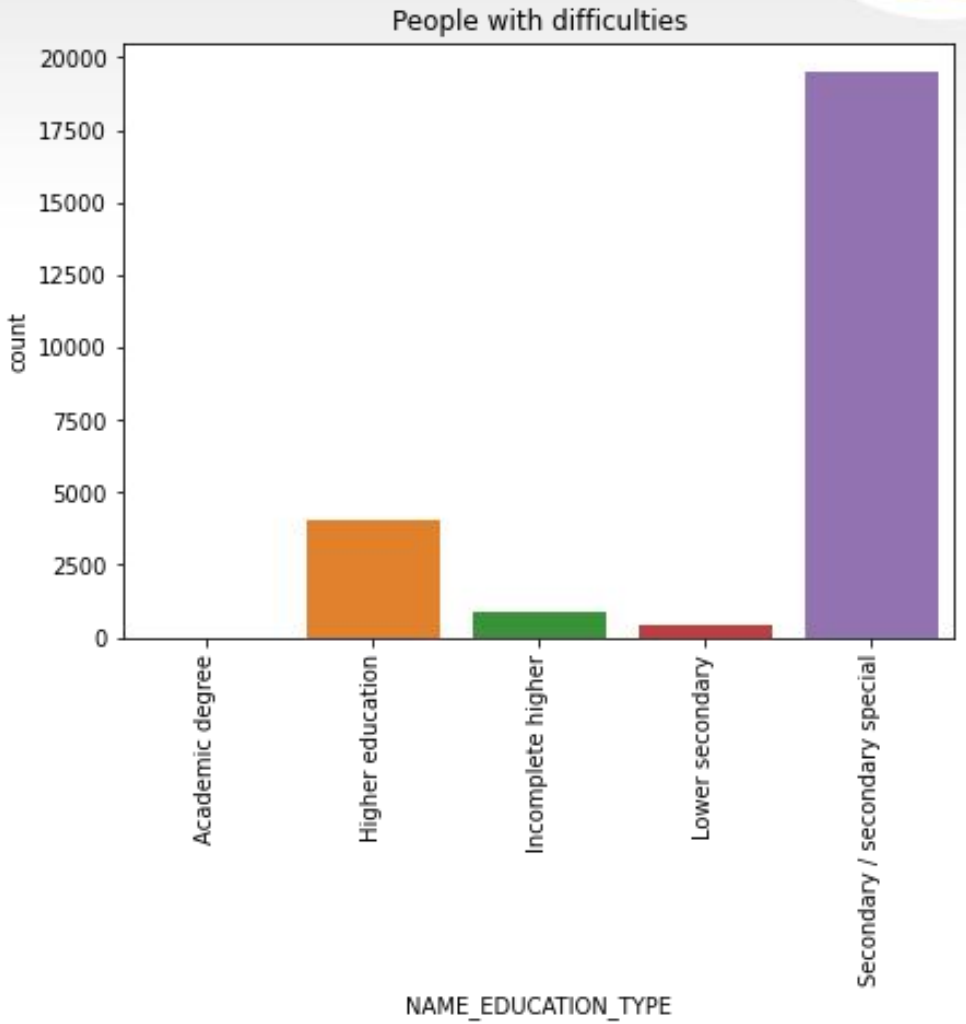




We see that Working people are the most in both category.
While the number of other variables are less in difficulty than not in difficulty.

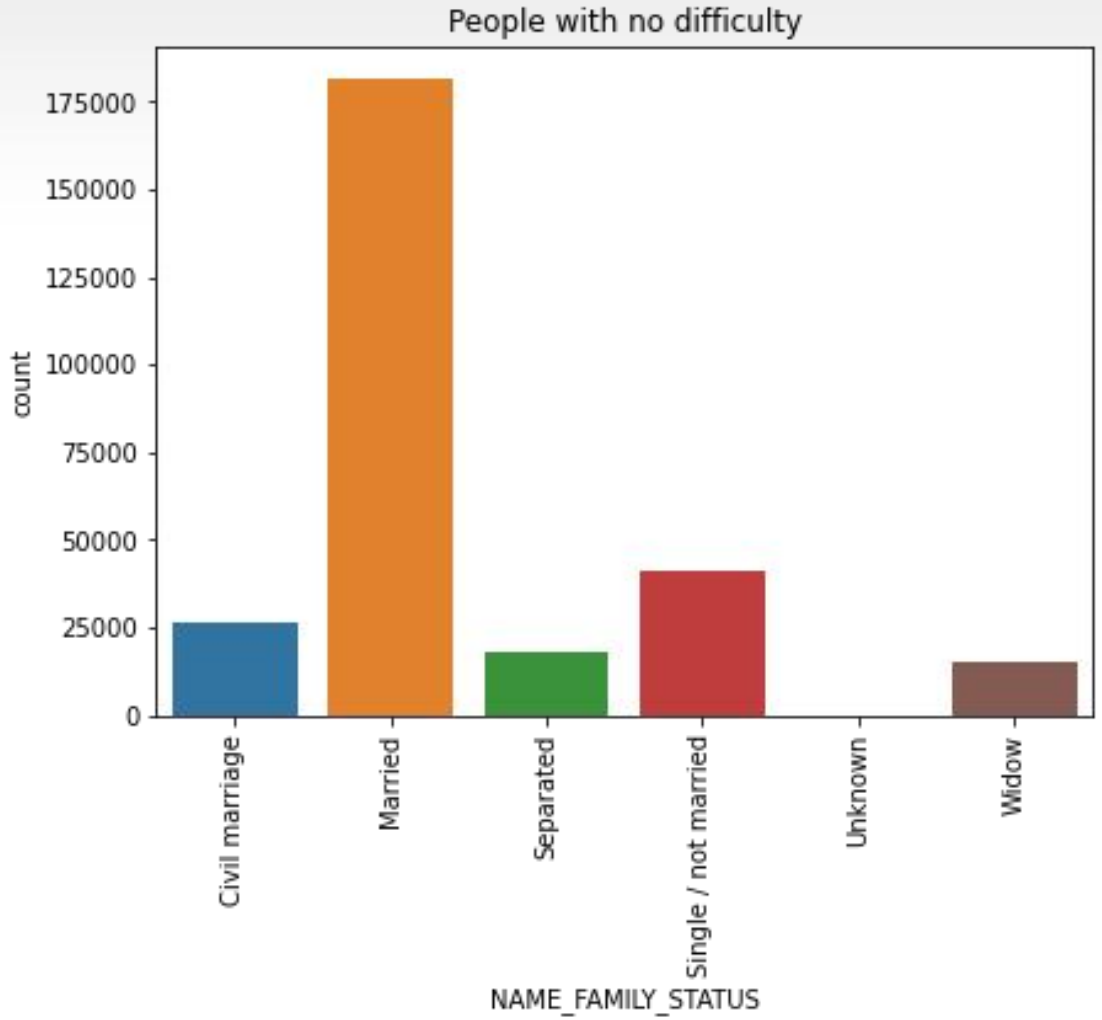
Education

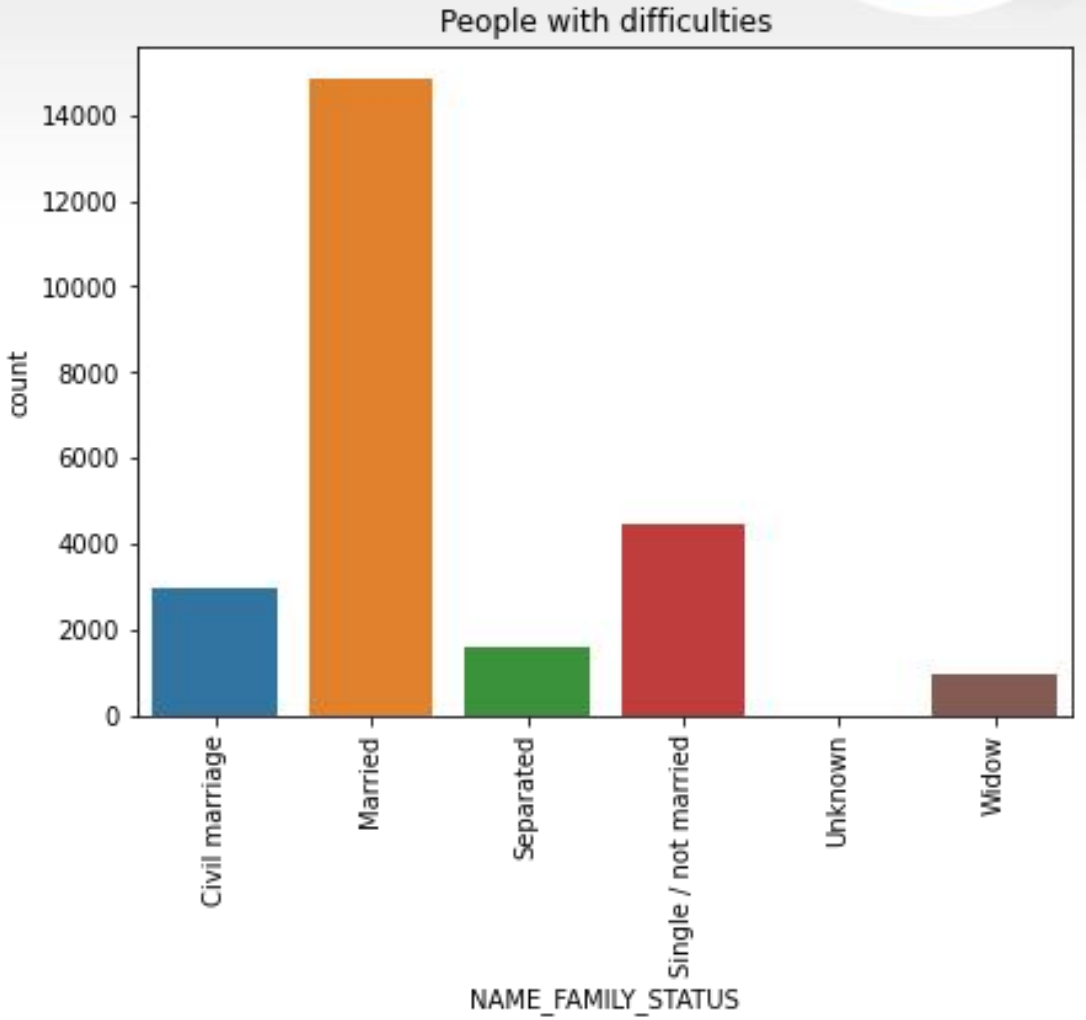




We can see that most of the people are educated to Secondary/secondary special.
Higher education are less in people with difficulty.

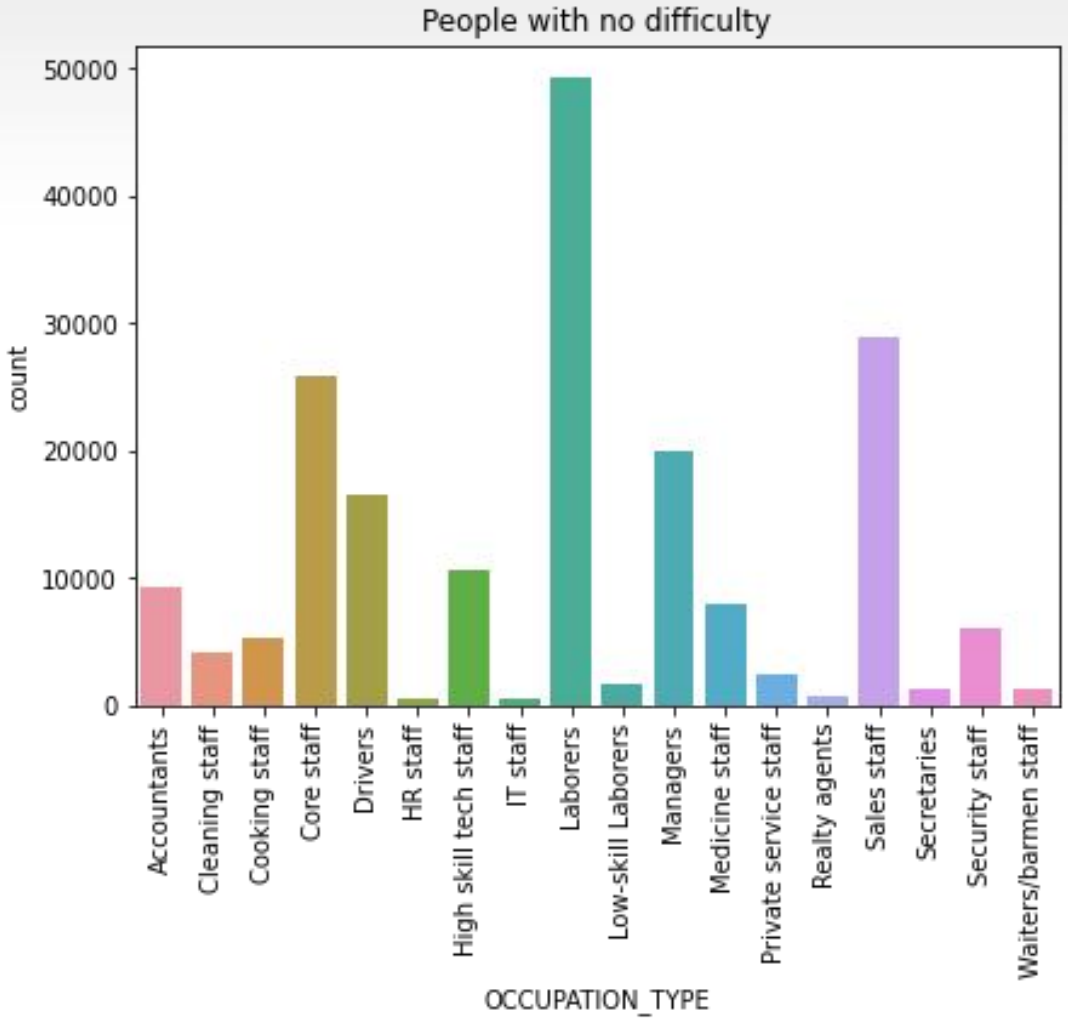
Family Status

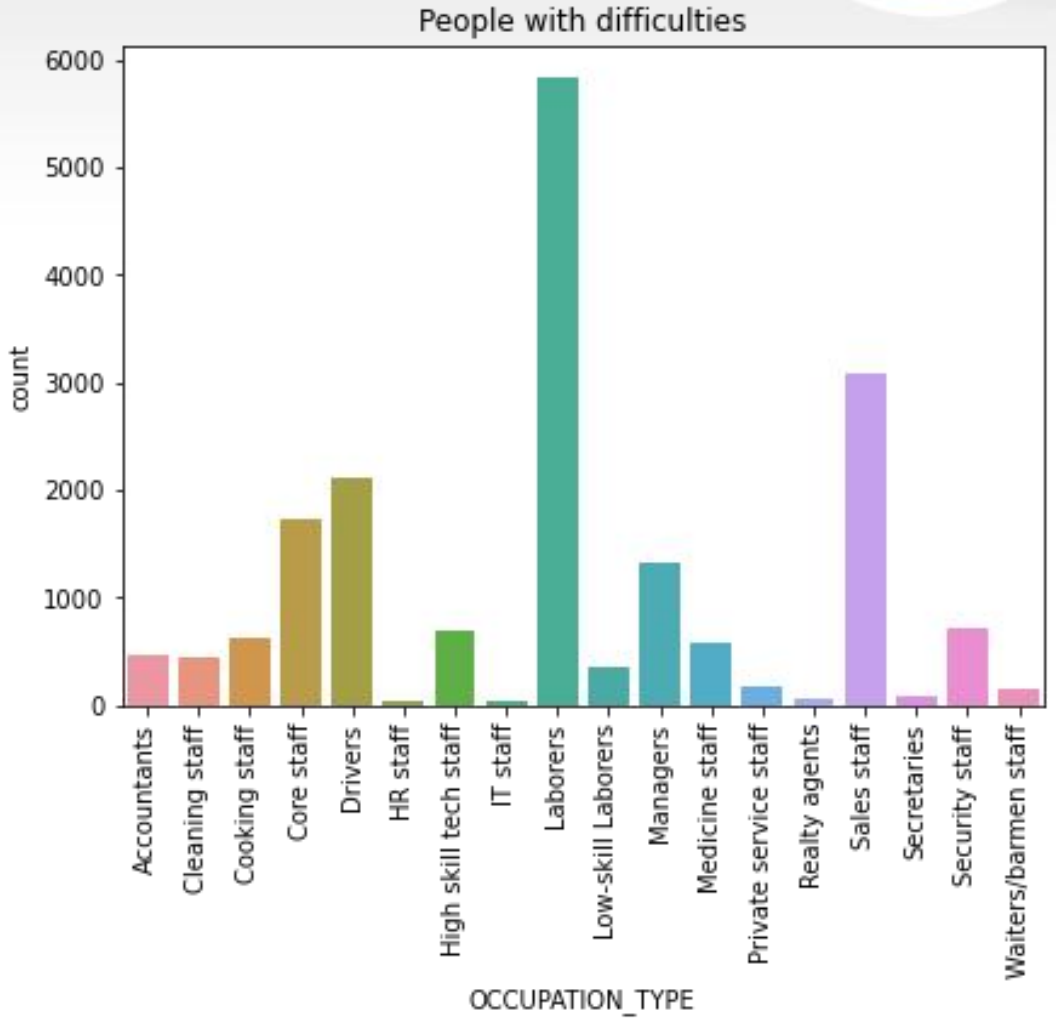




Here we see that most of the people are married. The bar of Single/not married is higher in people with difficulty. That means single people are more likely to default.

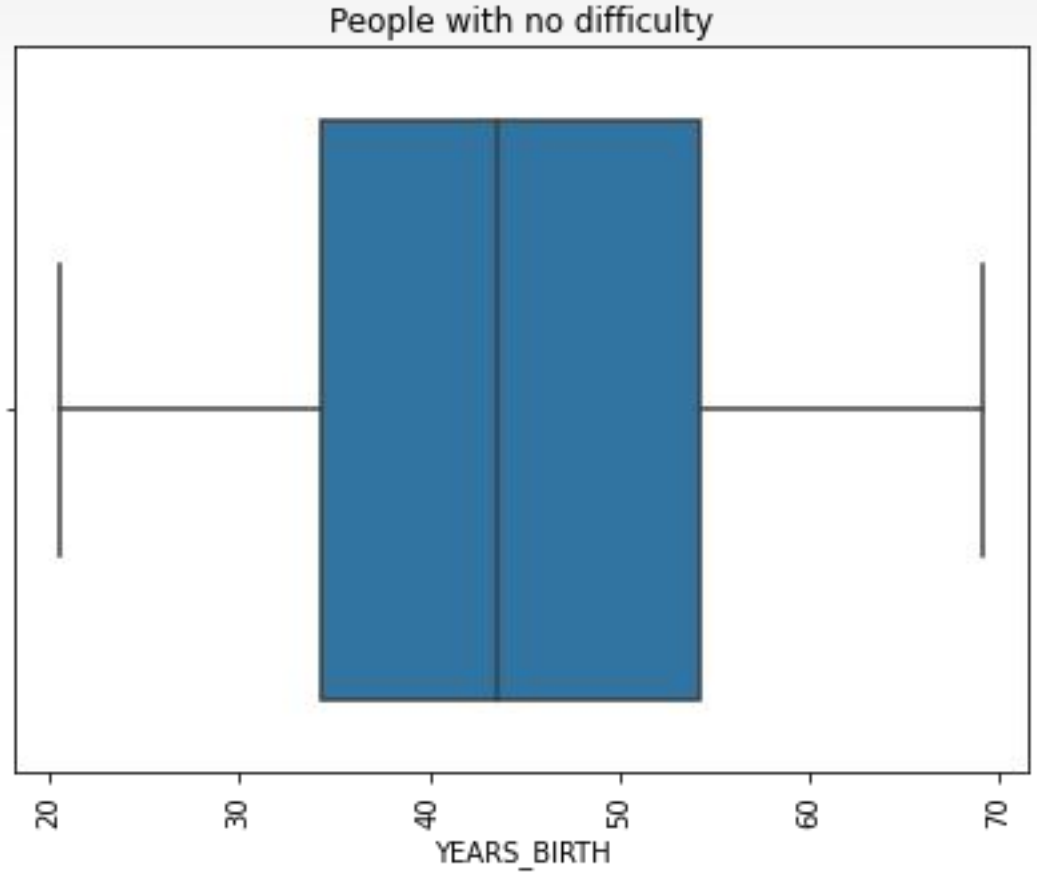
Occupation

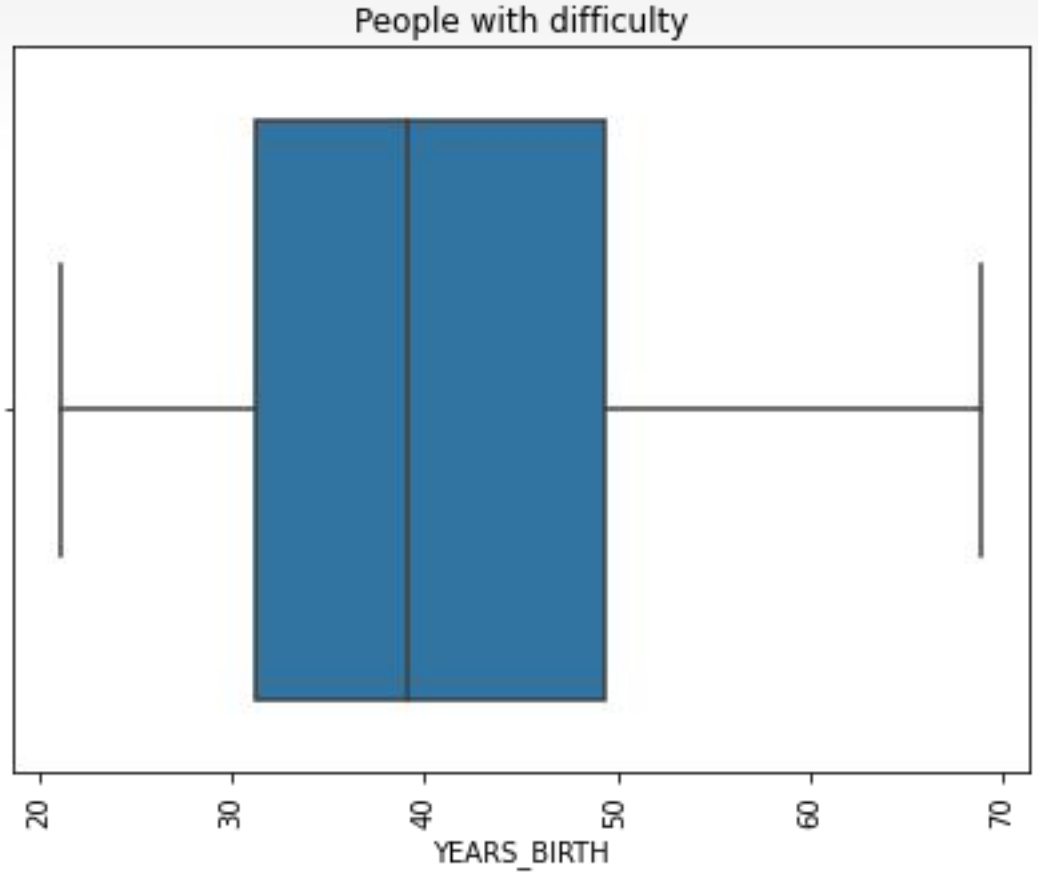




Here we see that Laborers are the most defaulting ones. The Sales staff and Drivers are also a risk.
Managers, Core staff and High skill tech staff are better at paying.

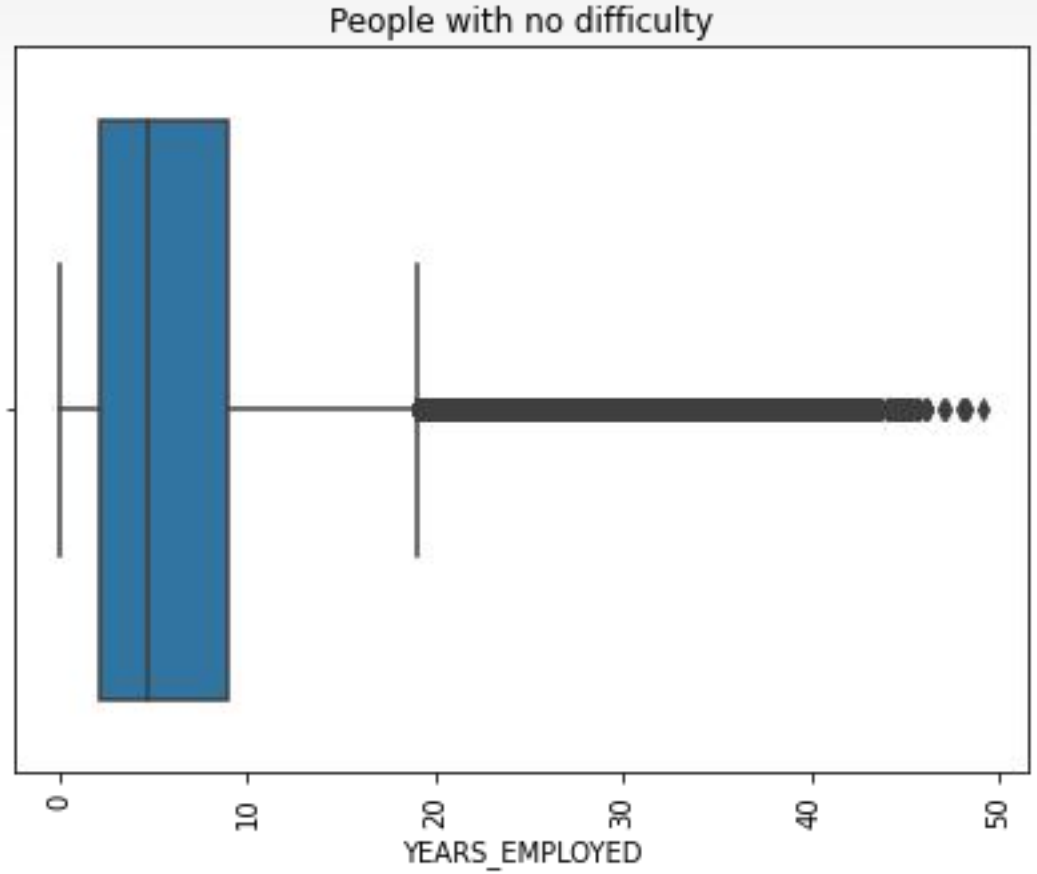
Age

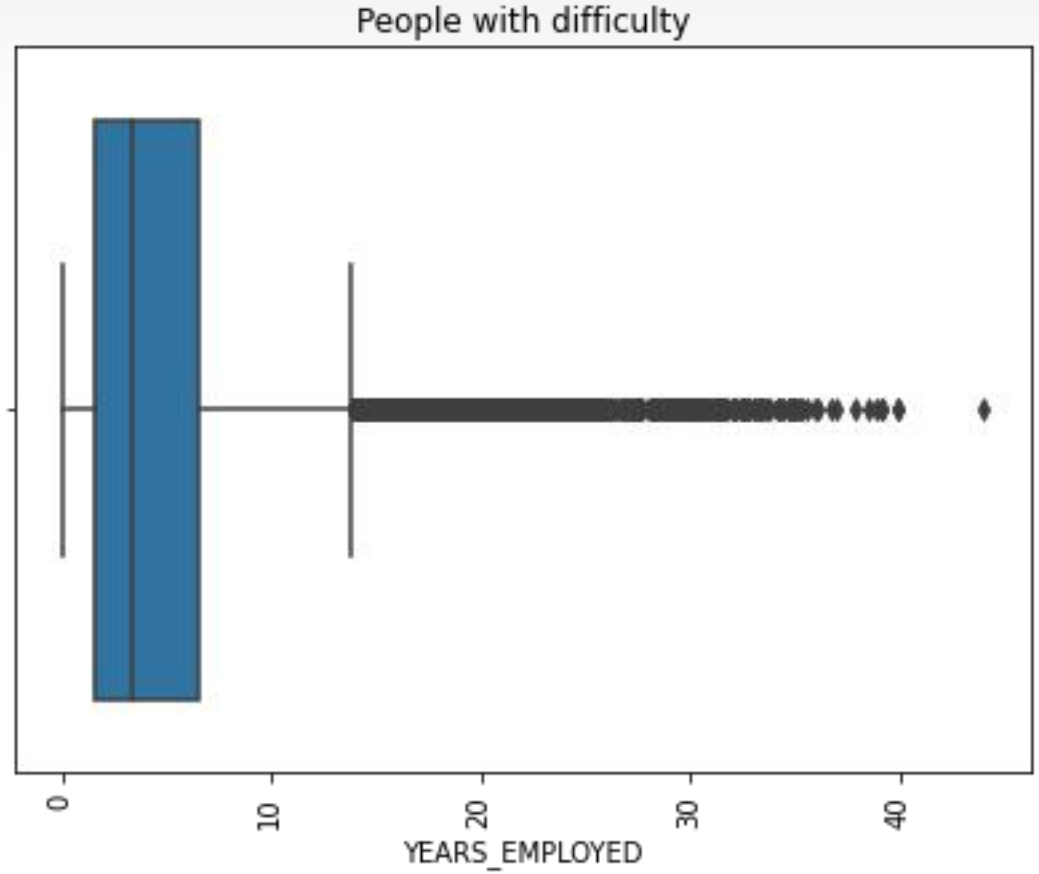




* We see here that values are same at min and max. But in q1 and q3, we see that people with difficulty are in 30-50 range.

YEARS EMPLOYED





We see that the upper fence for people with diificulty is less(around 15).

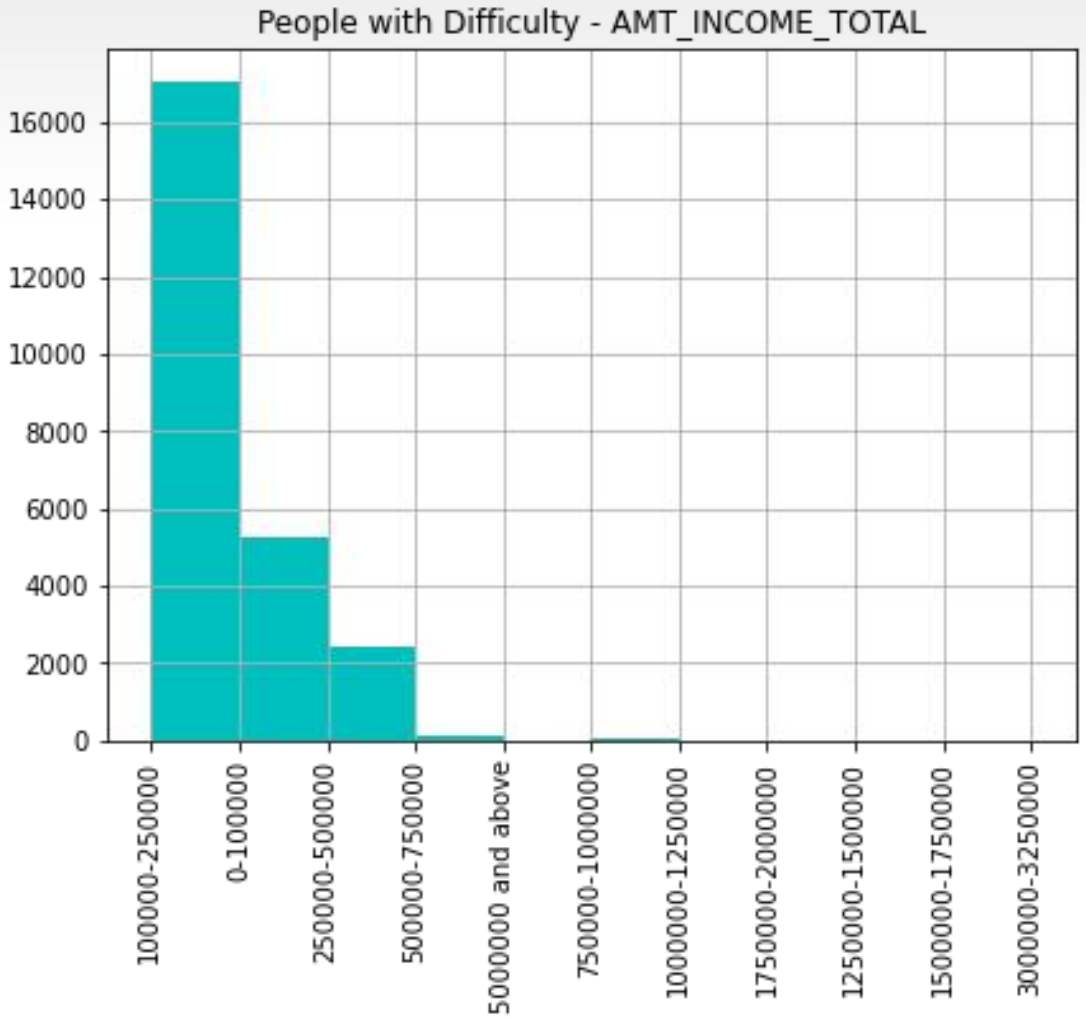
Top 10 Correlation for People with no difficulty

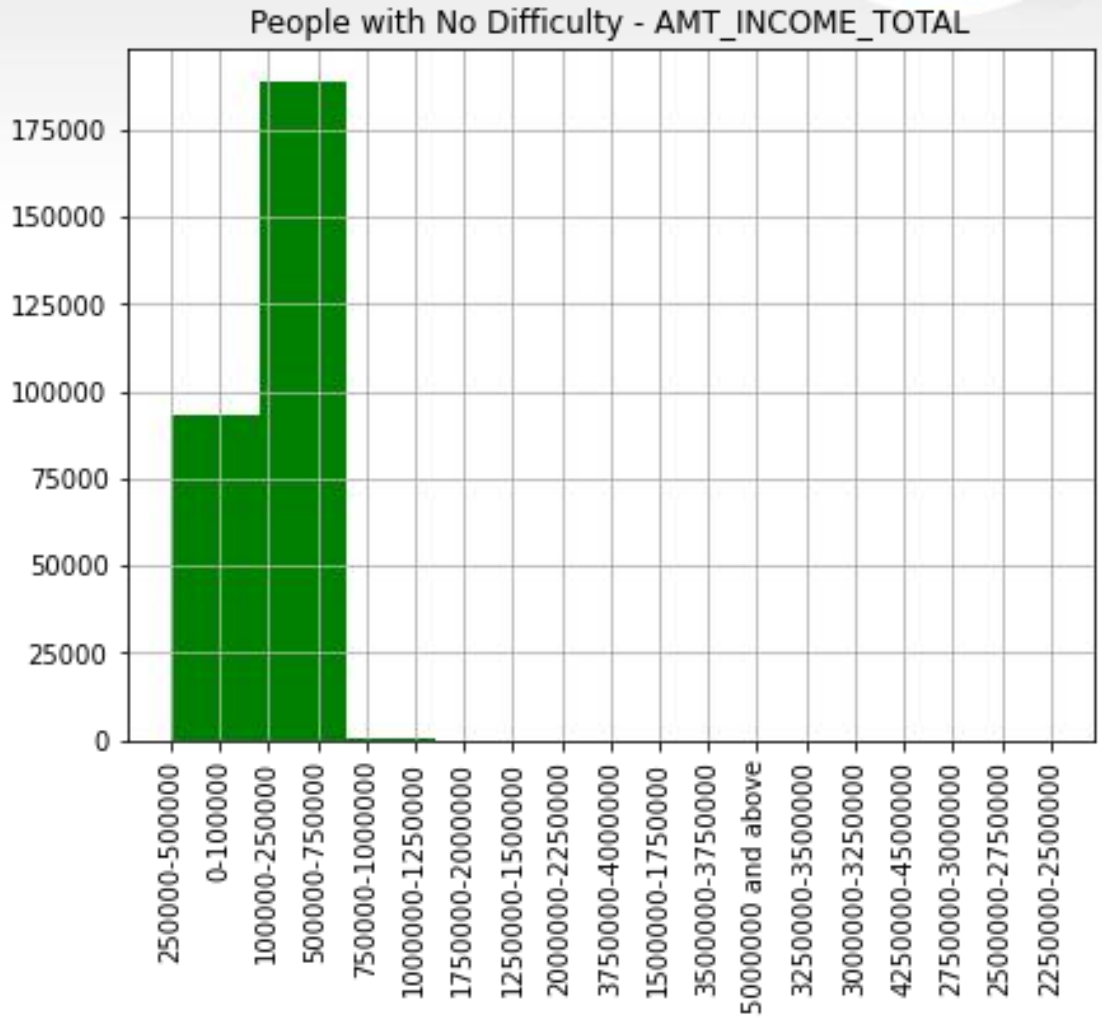
•	1	AMT_GOODS_PRICE	AMT_CREDIT	0.987250	0.987250
•	2	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686	0.776686
•	3	AMT_ANNUITY	AMT_CREDIT	0.771309	0.771309
•	4	YEARS_EMPLOYED	YEARS_BIRTH	0.626114	0.626114
•	5	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953	0.418953
•	6	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349462	0.349462
•	7	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799	0.342799
•	8	YEARS_REGISTRATION	YEARS_BIRTH	0.333150	0.333150
•	9	YEARS_ID_PUBLISH	YEARS_EMPLOYED	0.276664	0.276664
•	10	YEARS_ID_PUBLISH	YEARS_BIRTH	0.271315	0.271315

Top 10 Correlation for People who default

•	1	AMT_GOODS_PRICE	AMT_CREDIT	0.983103	0.983103
•	2	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699	0.752699
•	3	AMT_ANNUITY	AMT_CREDIT	0.752195	0.752195
•	4	YEARS_EMPLOYED	YEARS_BIRTH	0.582187	0.582187
•	5	YEARS_REGISTRATION	YEARS_BIRTH	0.289111	0.289111
•	6	YEARS_ID_PUBLISH	YEARS_BIRTH	0.252867	0.252867
•	7	YEARS_ID_PUBLISH	YEARS_EMPLOYED	0.229094	0.229094
•	8	YEARS_REGISTRATION	YEARS_EMPLOYED	0.192454	0.192454
•	9	YEARS_BIRTH	EXT_SOURCE_3	0.171618	0.171618
•	10	EXT_SOURCE_2	REGION_POPULATION_RELATIVE	0.169751	0.169751

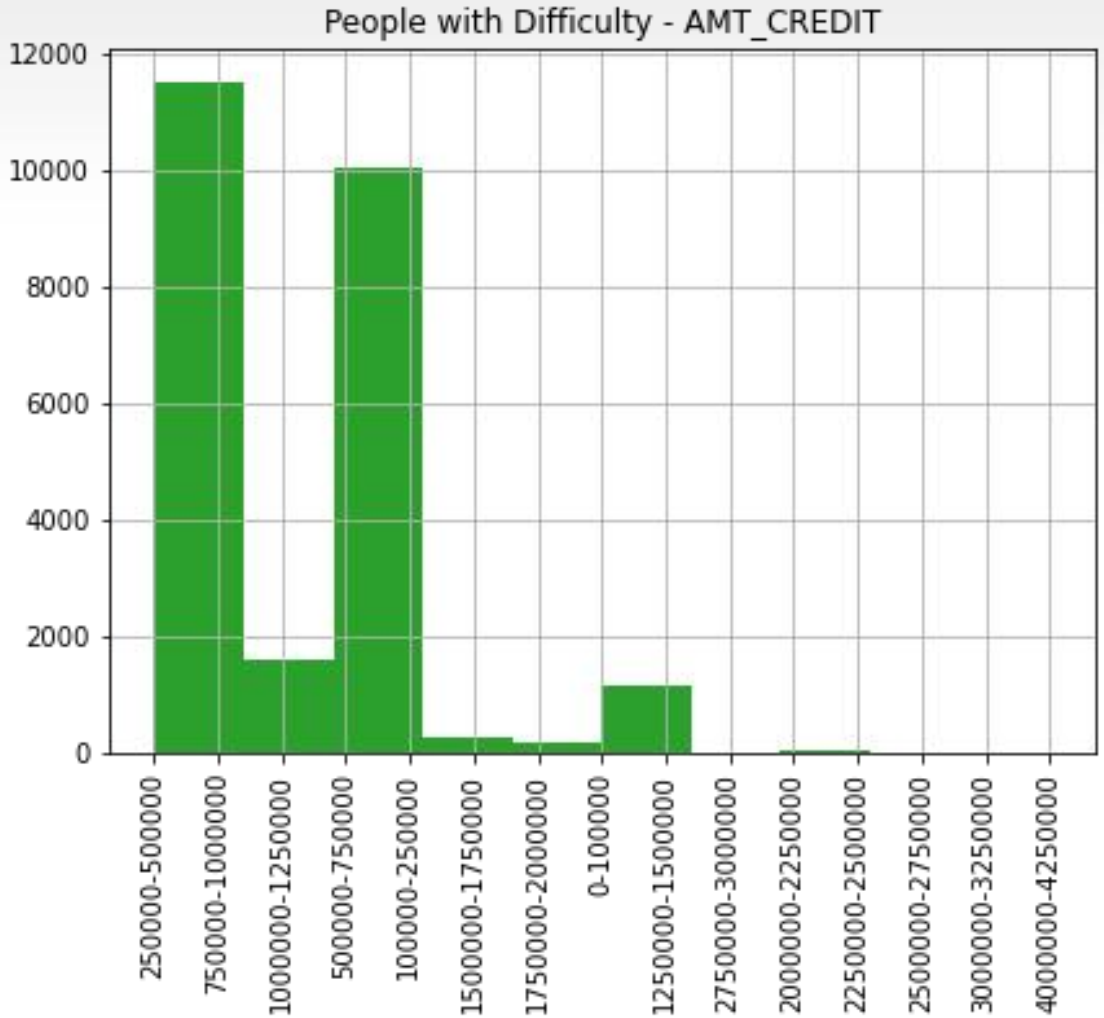
Total Income

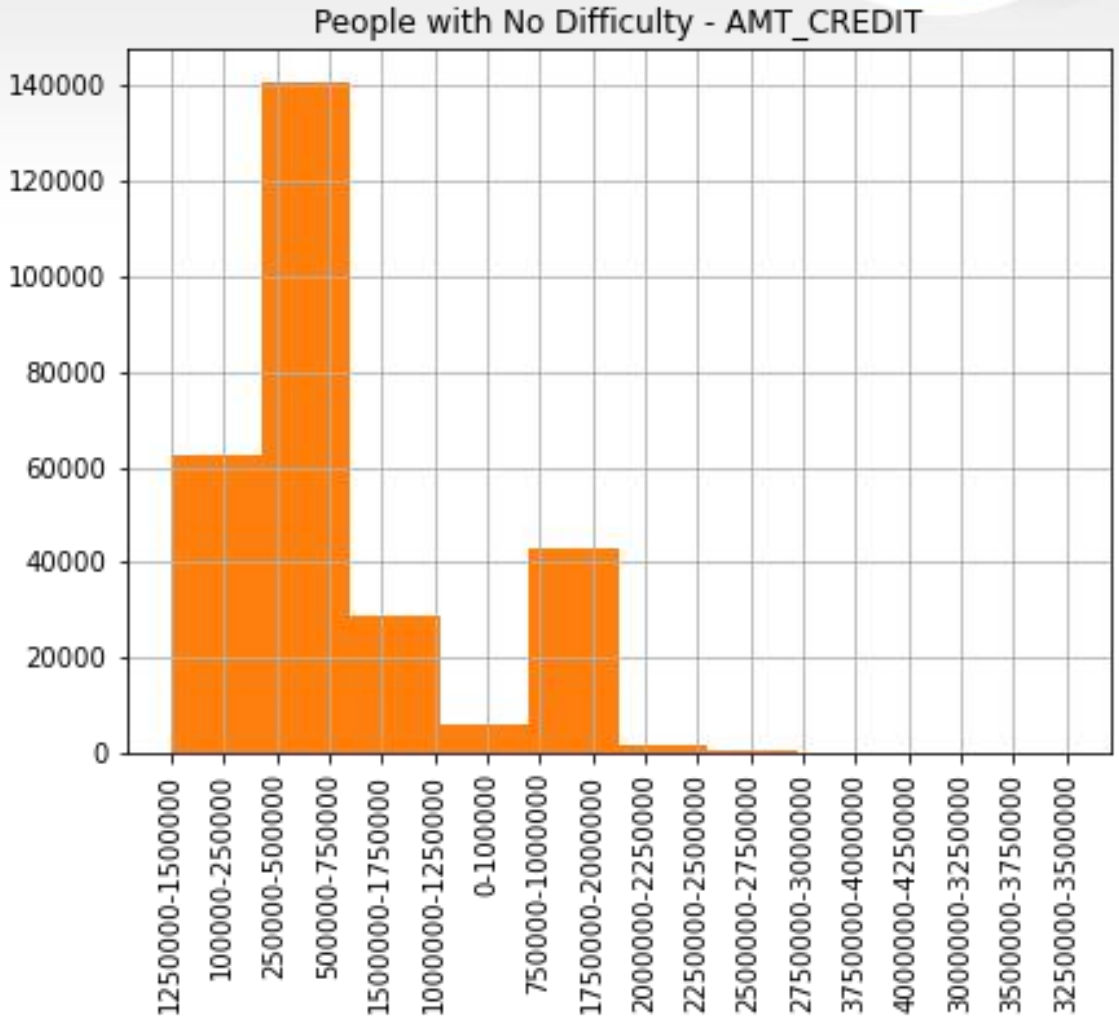




We see that most people with payment difficulties are around total income 0 TO 500000

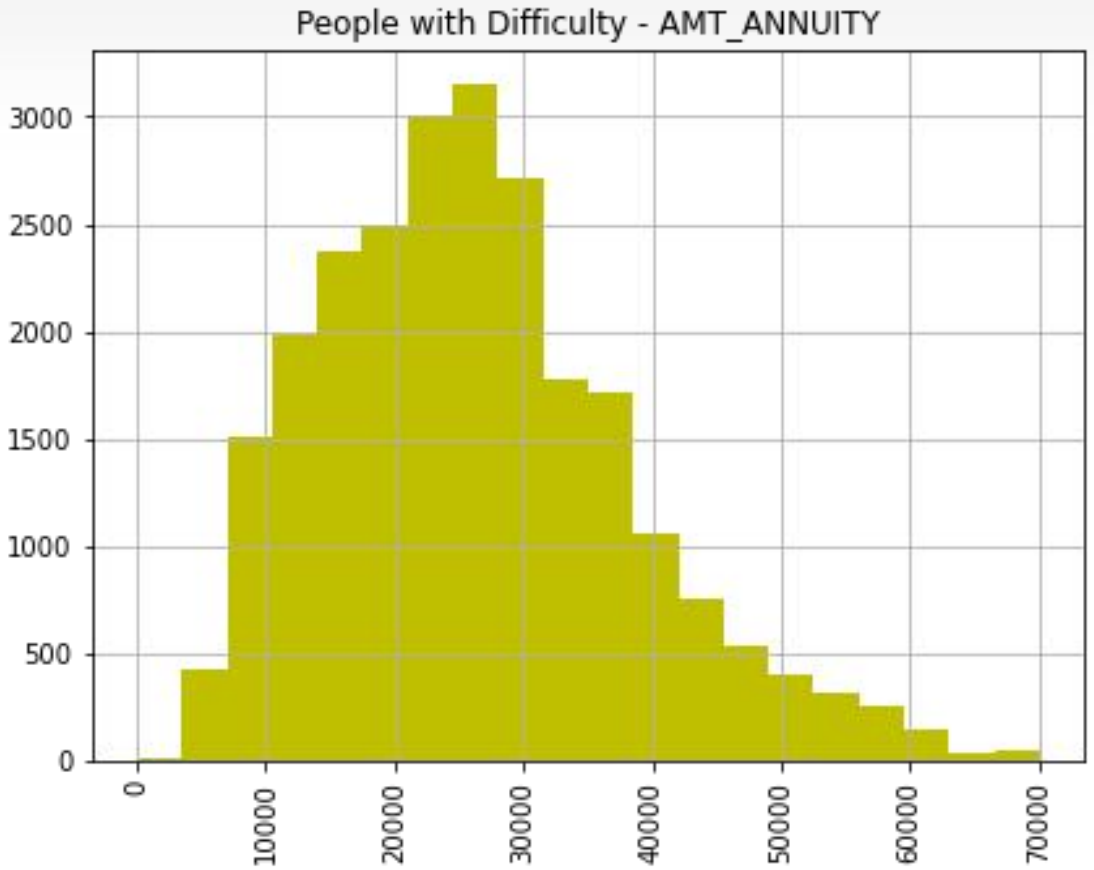
Credit

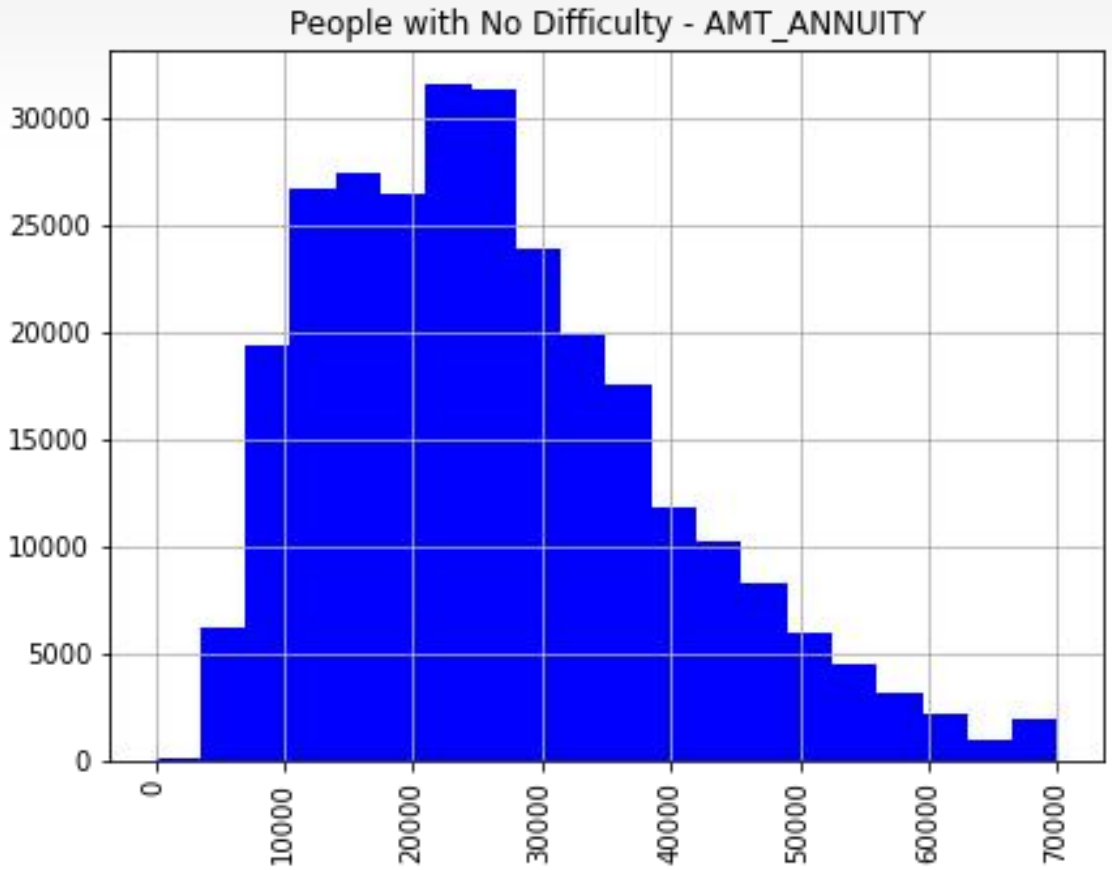




Most of the loan default is for Clients whose Credit is between 100000 TO 1100000 Loan Amounts.

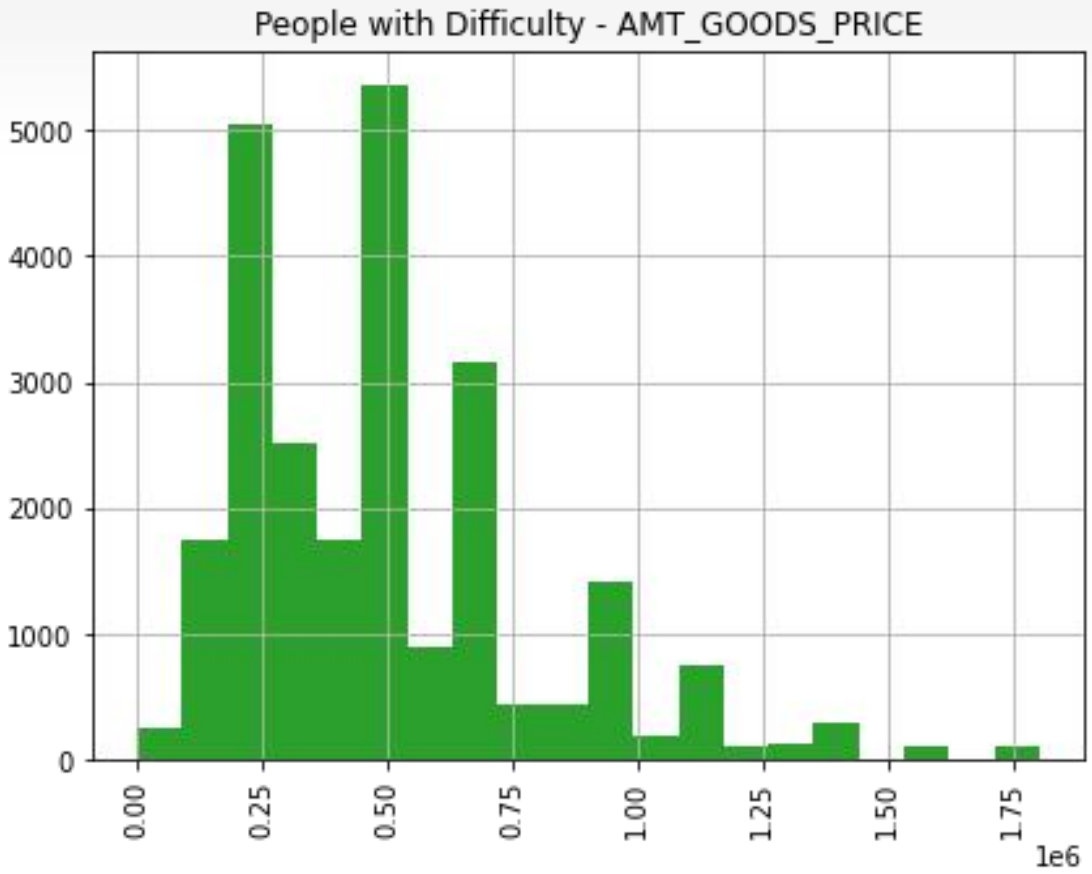
Annuity

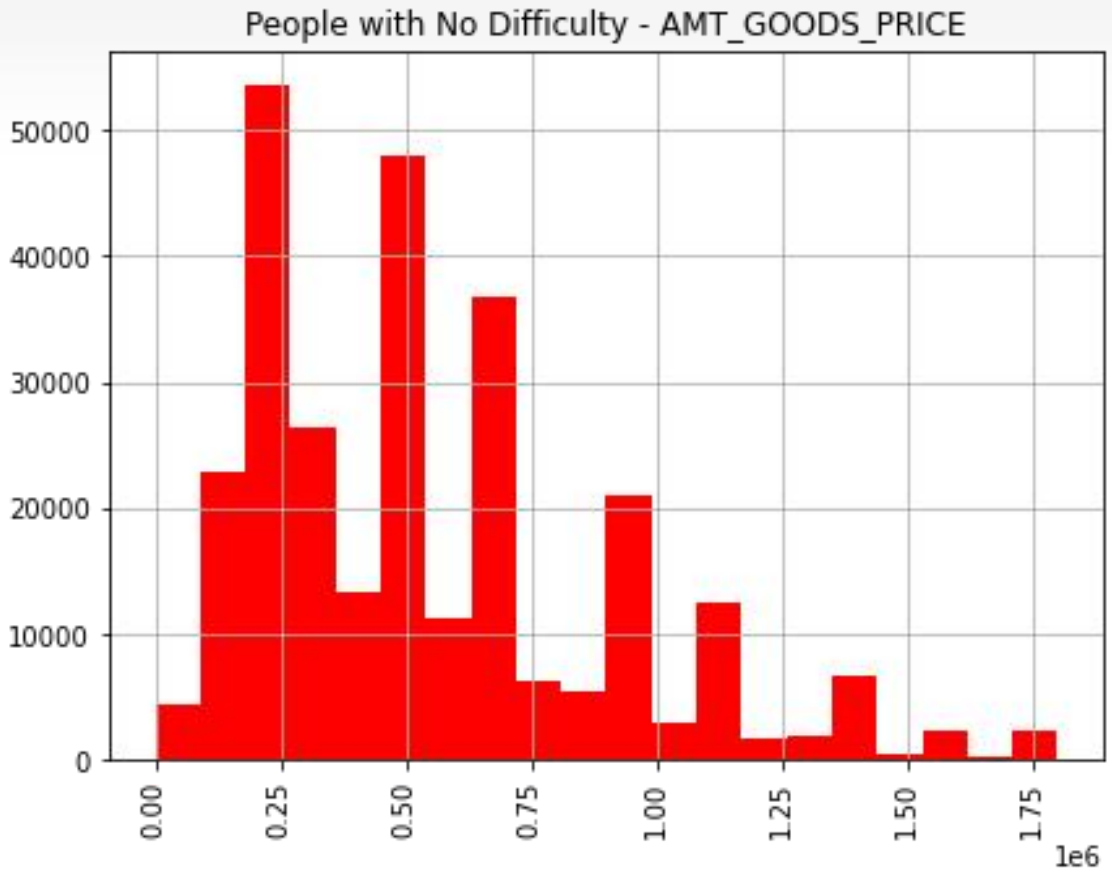




We see that highest no. of people with payment difficulties are people whose annuity is between 9000 TO 38000.

Goods Price

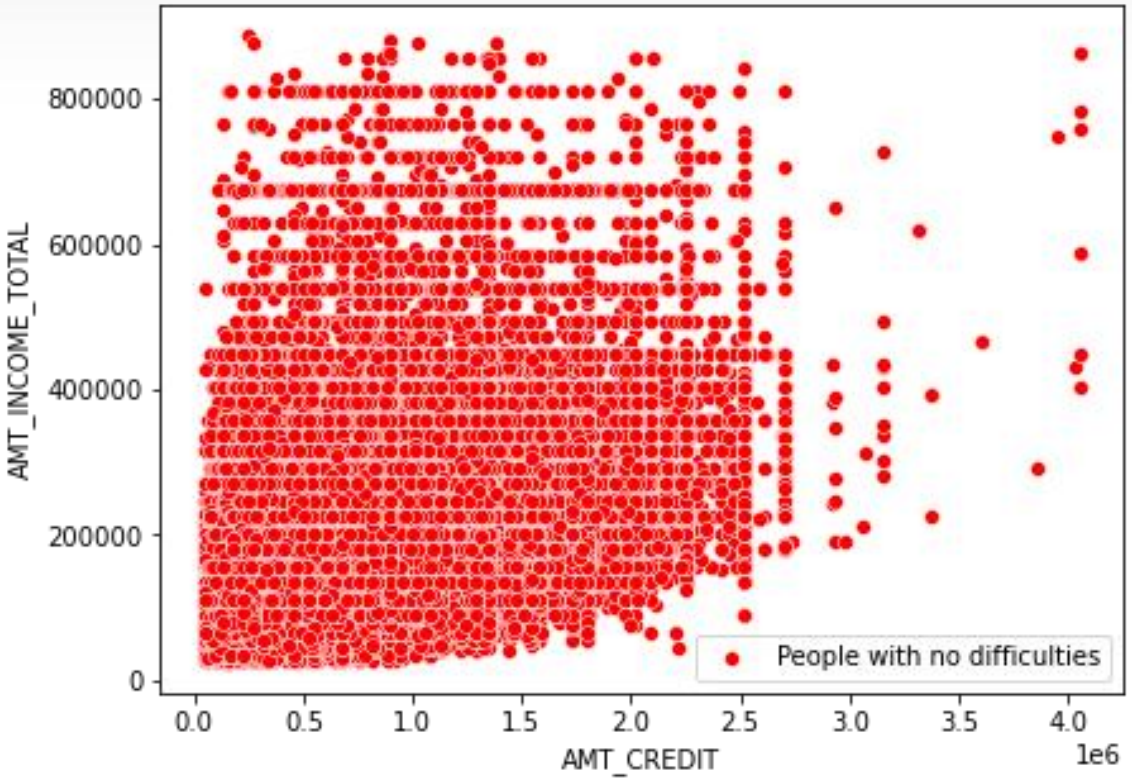


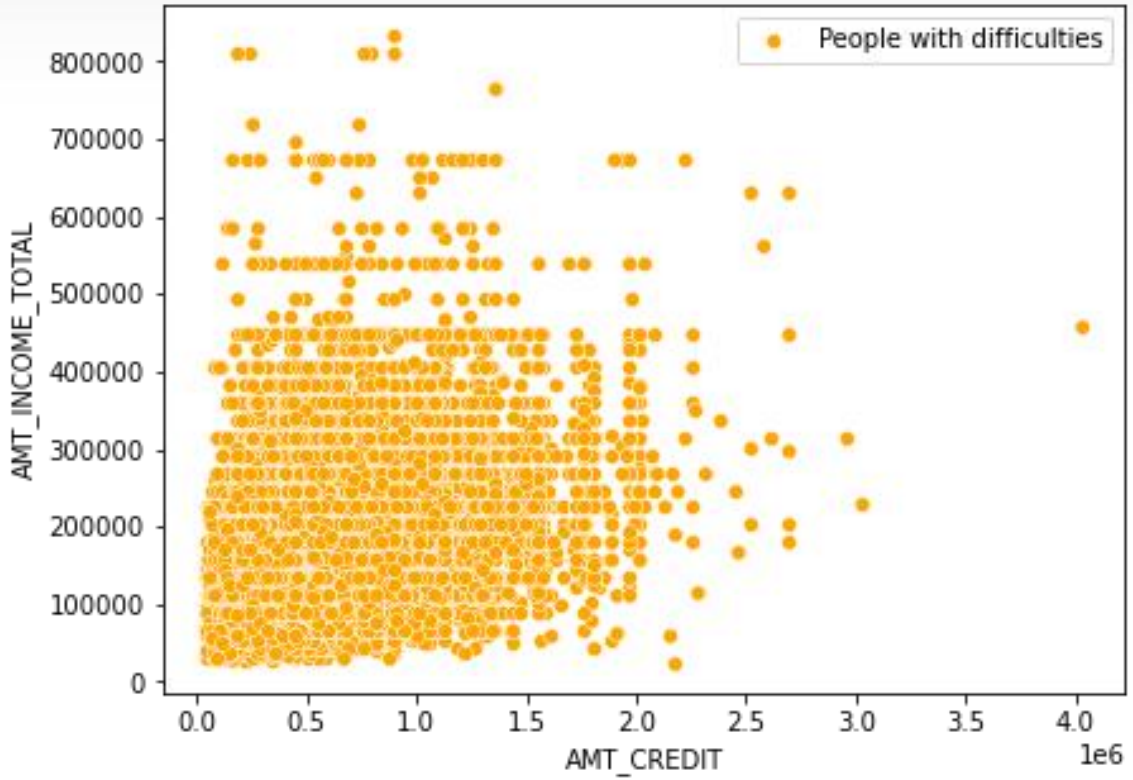


We see that highest no. of people with payment difficulties are people whose GOODS PRICE is between 100000 TO 750000.

BIVARIATE ANALYSIS

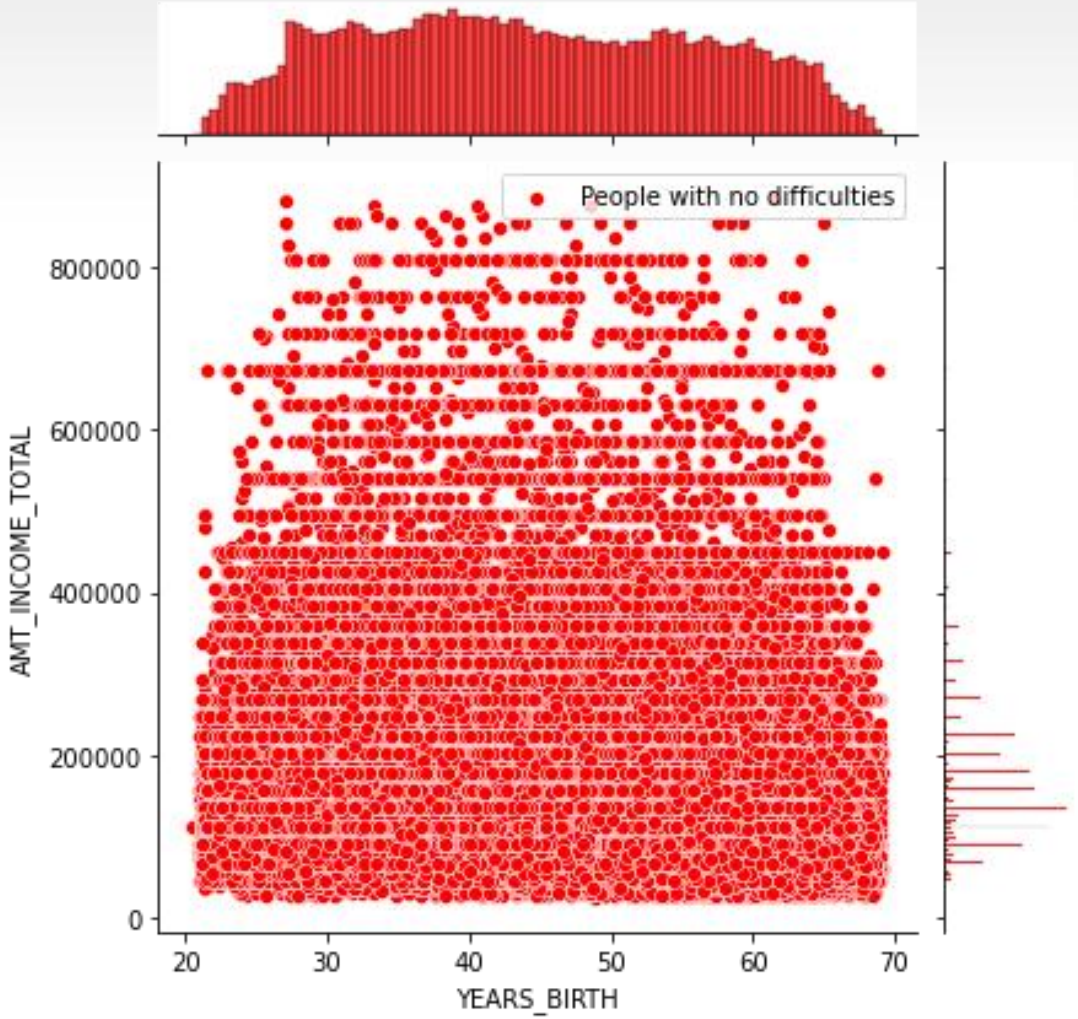
Total Income VS Credit

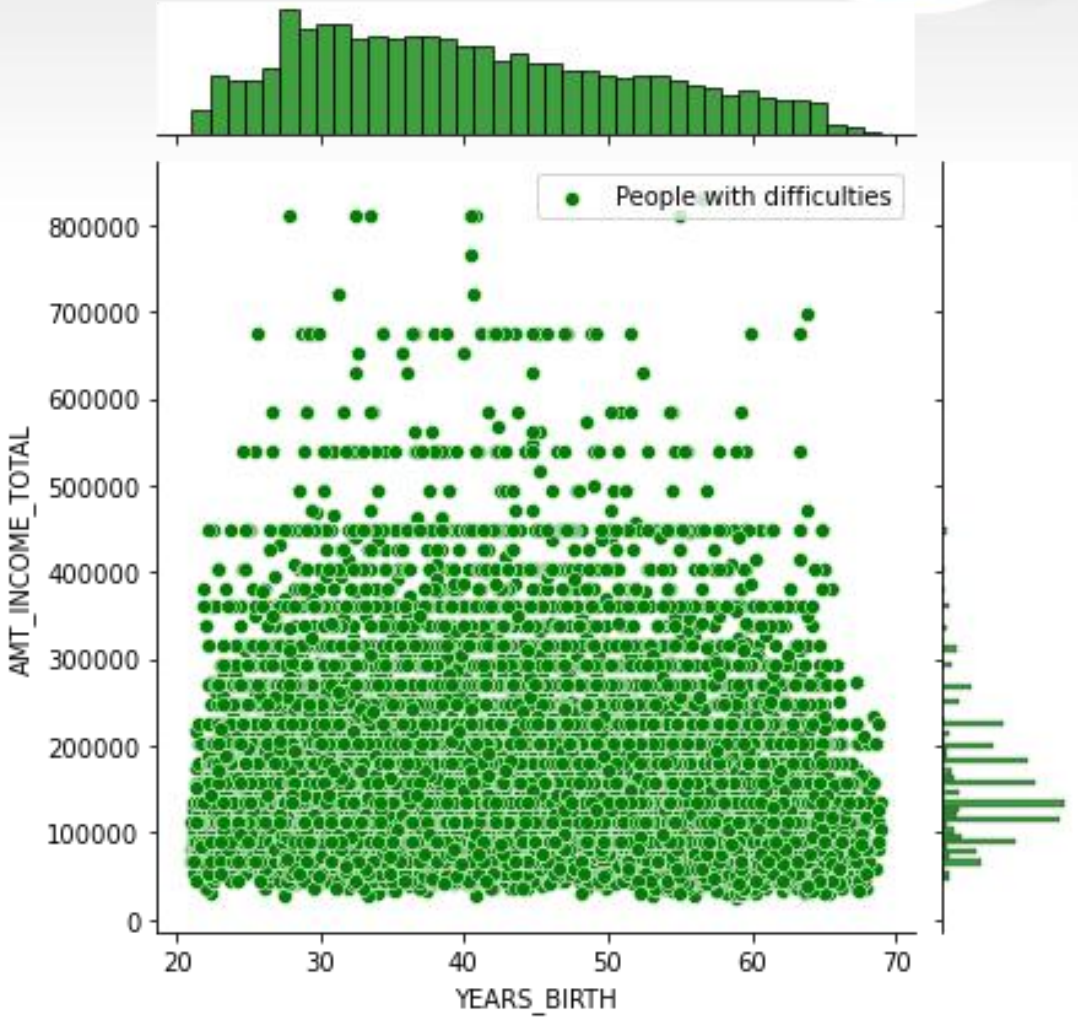




We see most of data is together. The people with difficulty are in the range of 0-1800000 credit. Income wise around 500000.

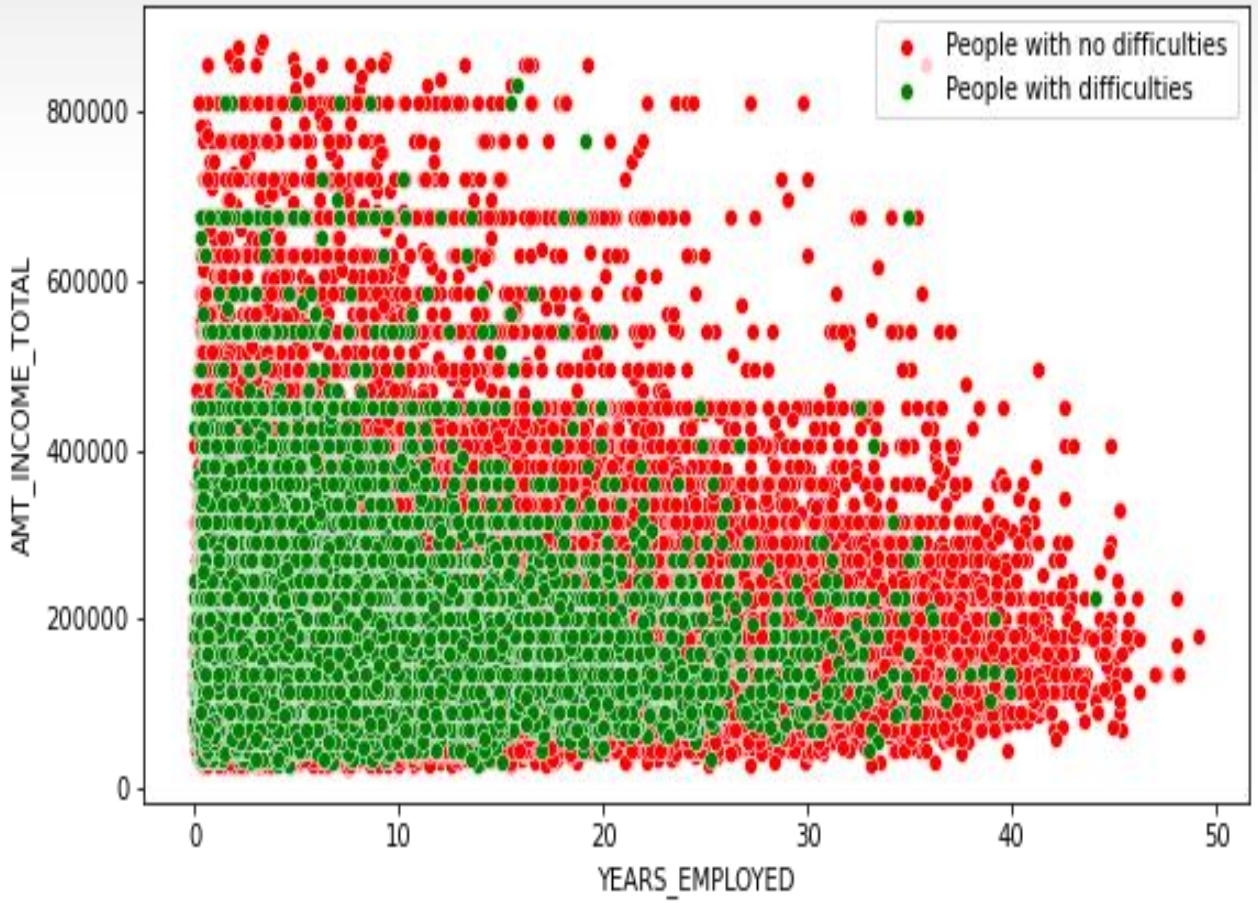
Income VS Age

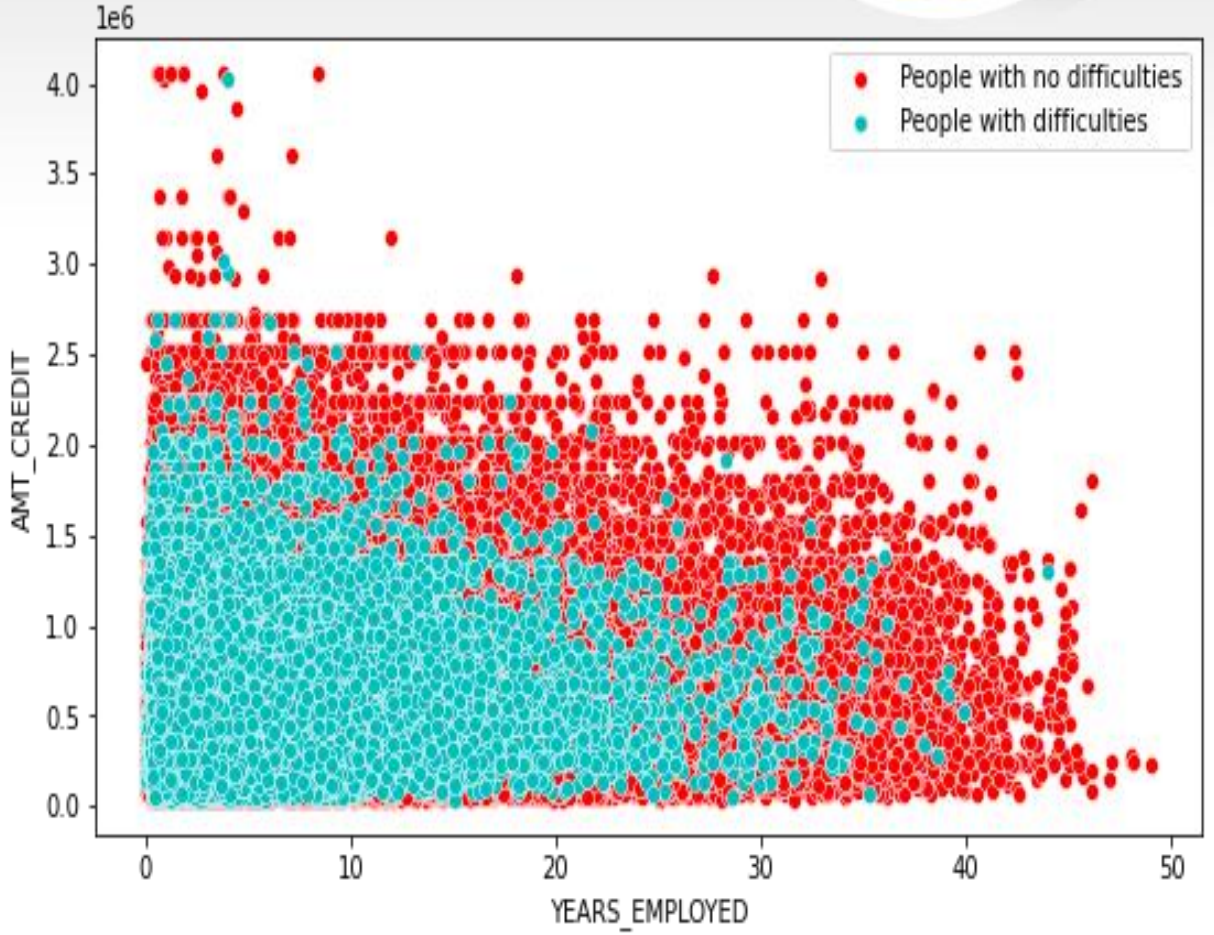




We see that people's income does not show much change with age but people with difficulties show a downward trend in the age section. That means no. of people with low age are more in risk of failing to pay installments when compared to older people.

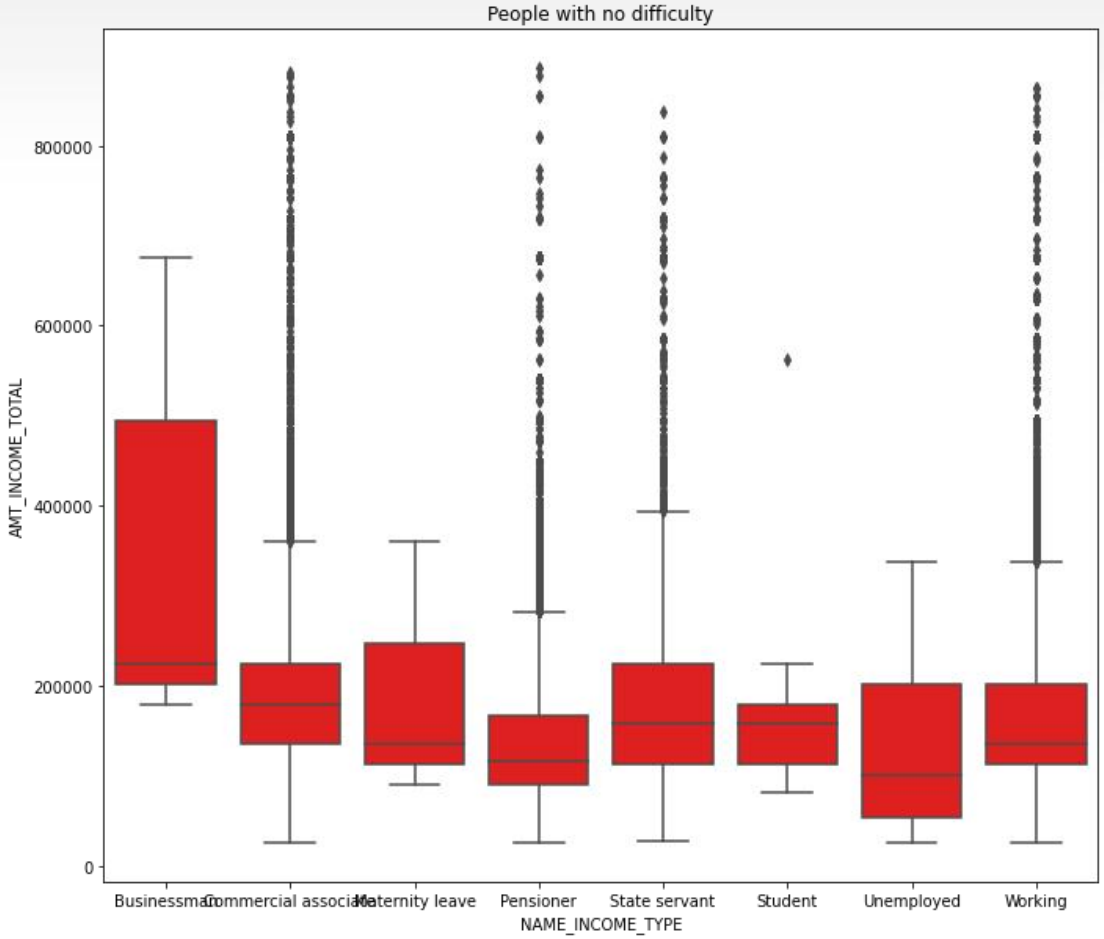
Income & Credit VS Years Employed

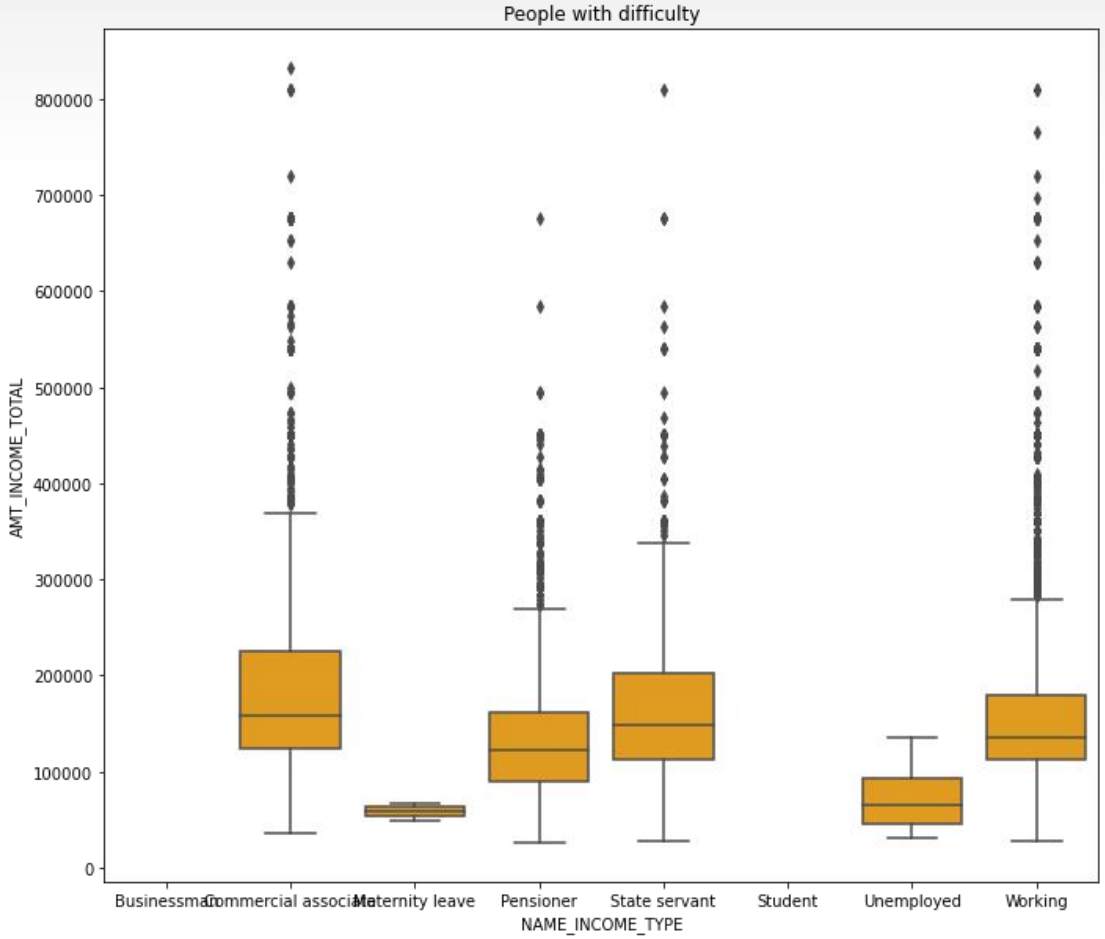




We see that both Income and Credit goes down with Years Employed.
But both Income and Credit as well as Years Employed are less in people with difficulty.
Around 30 years employed people are in difficulty.

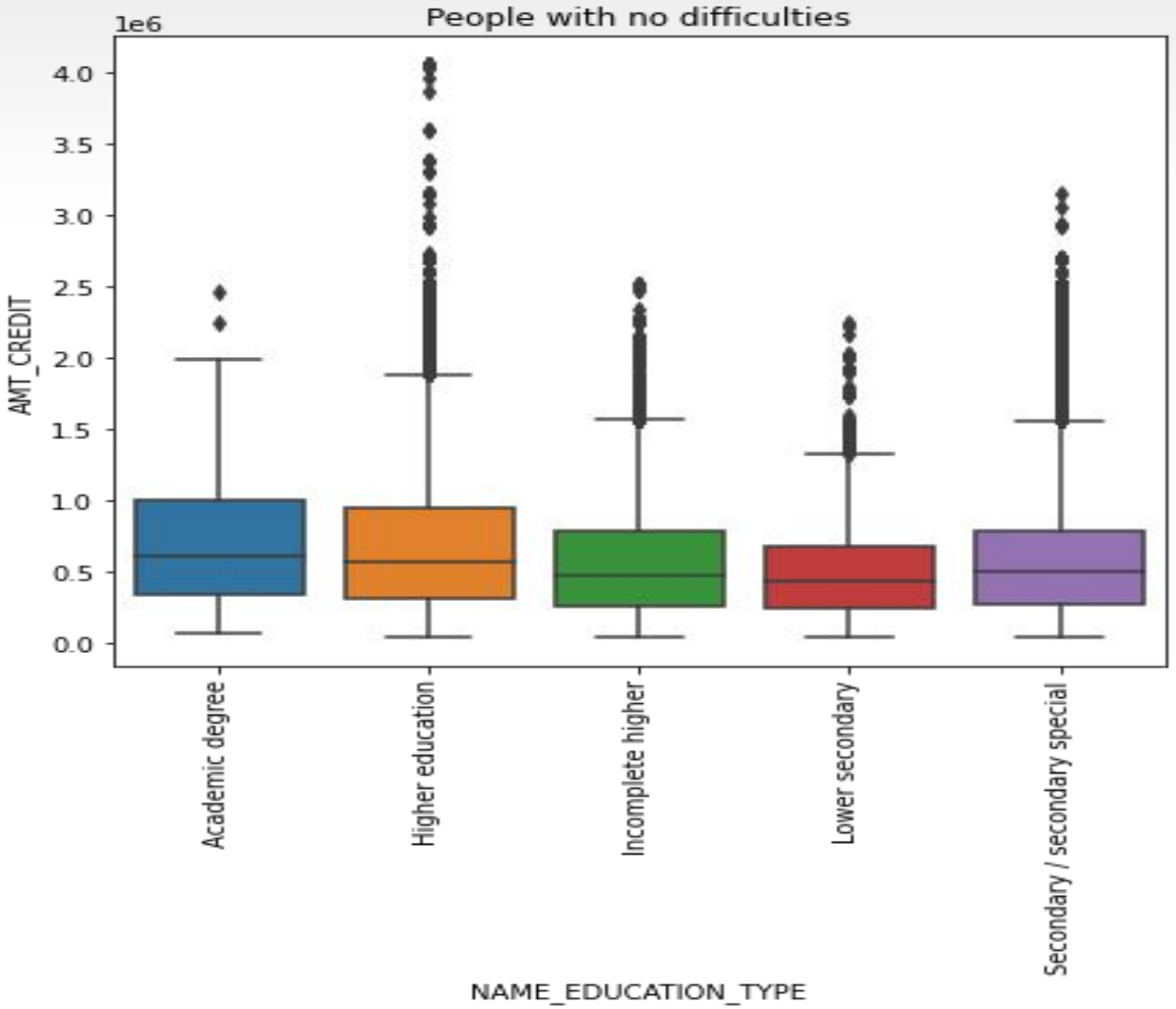
Total Income VS Income Type

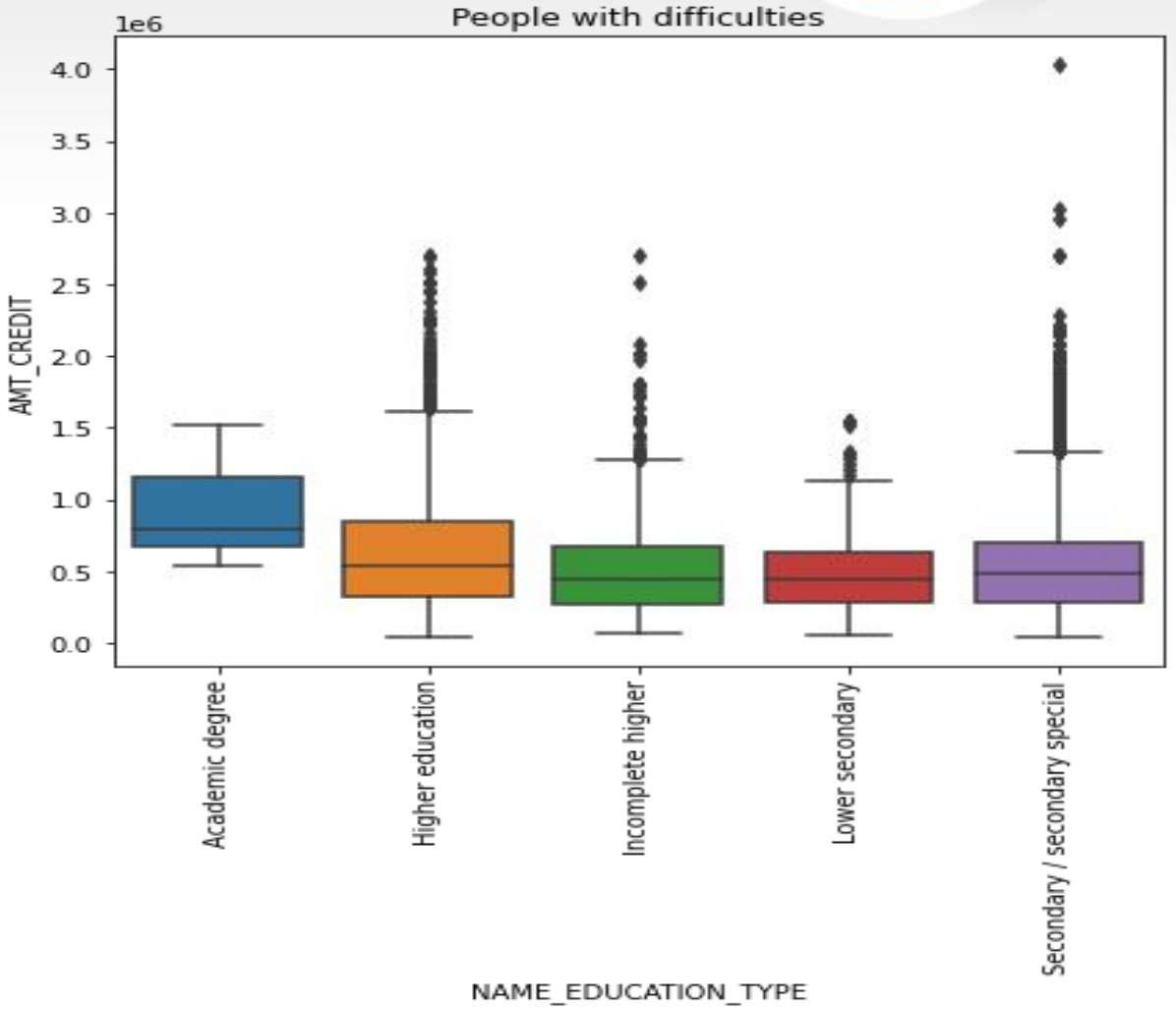




1. We see that businessman & Student do no to show defaults while maternity leave is less.
2. Pensioners & State servant with default are in Income range of 500000.
3. Unemployed with default are below 200000. Working People are the same in both category.

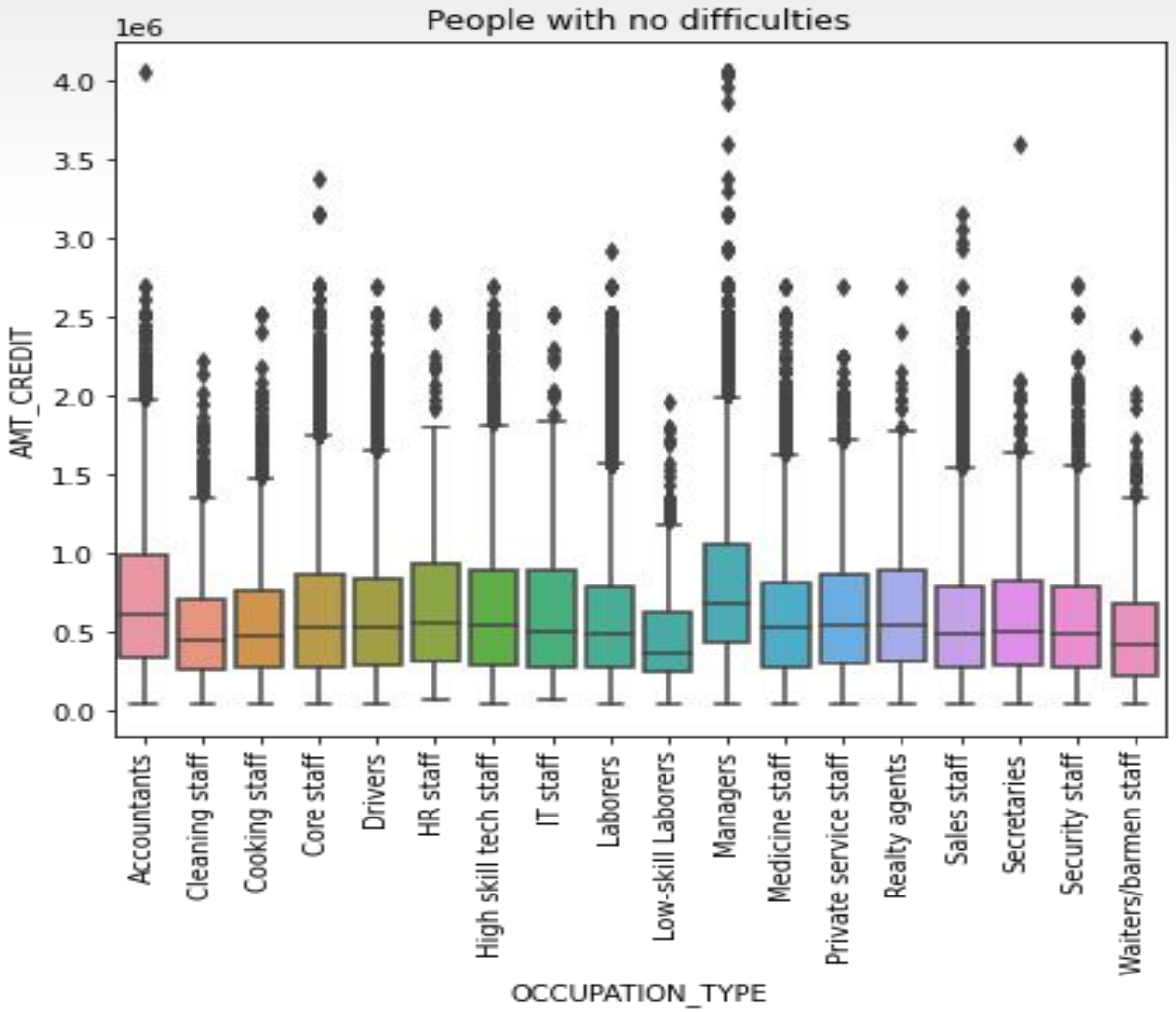
Credit VS Education

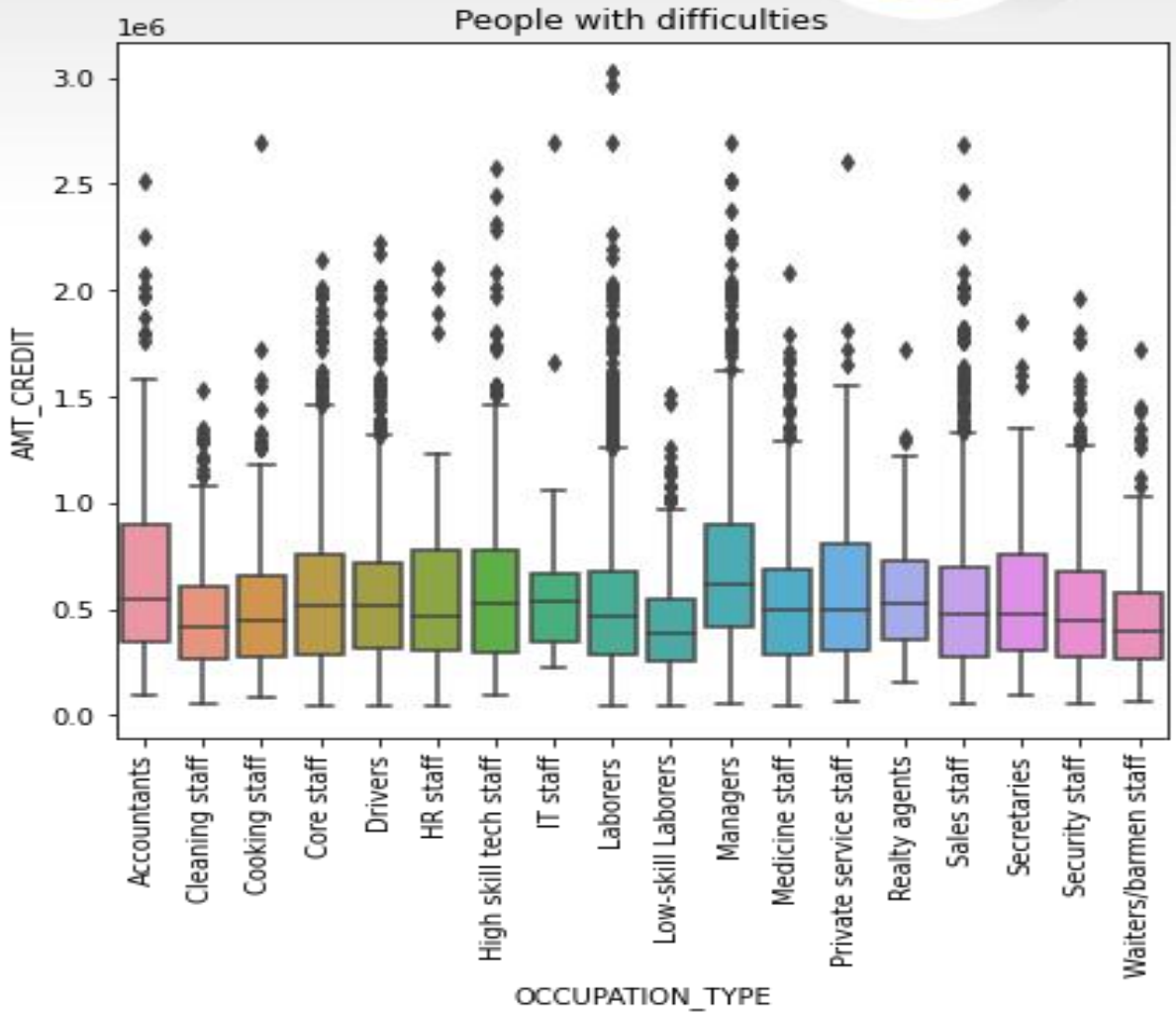




We see that the customers with diificulties are less in Academic degree at 500000 to 1500000.
Higher Education people in difficulty are around 2800000.

Credit VS Occupation

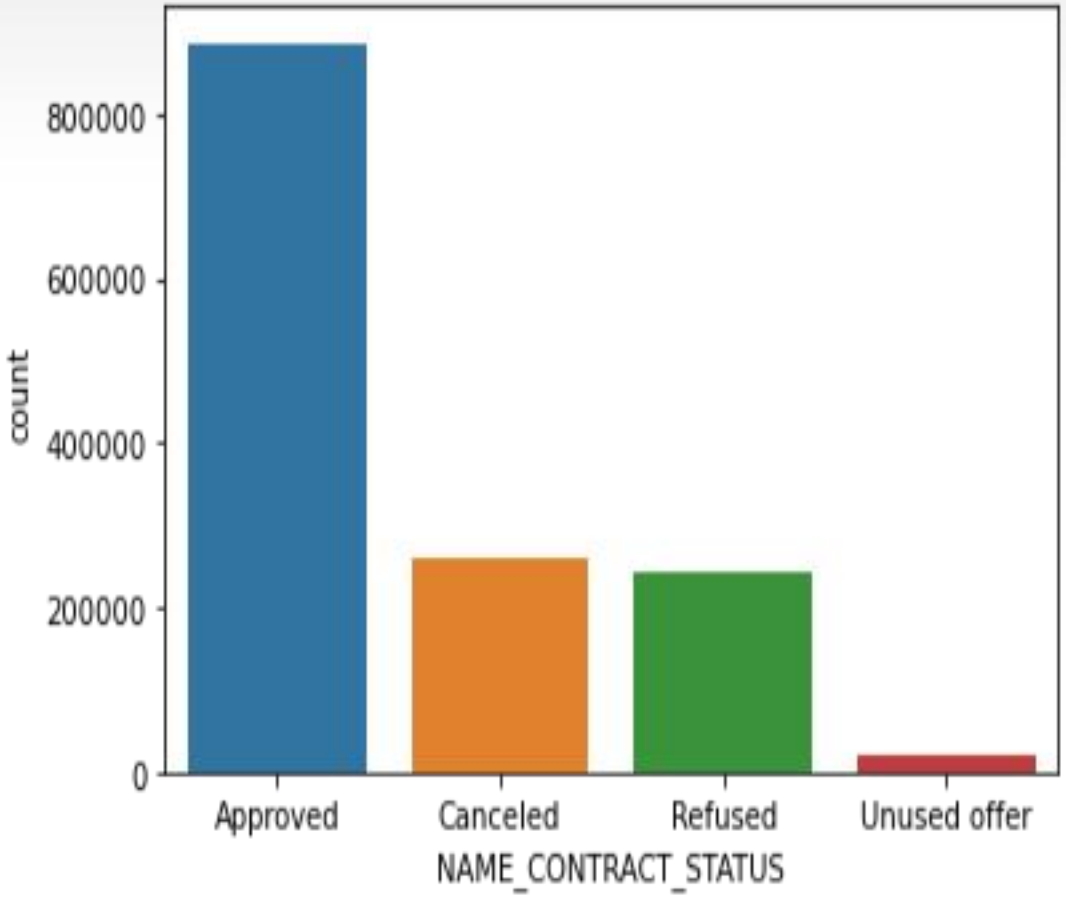




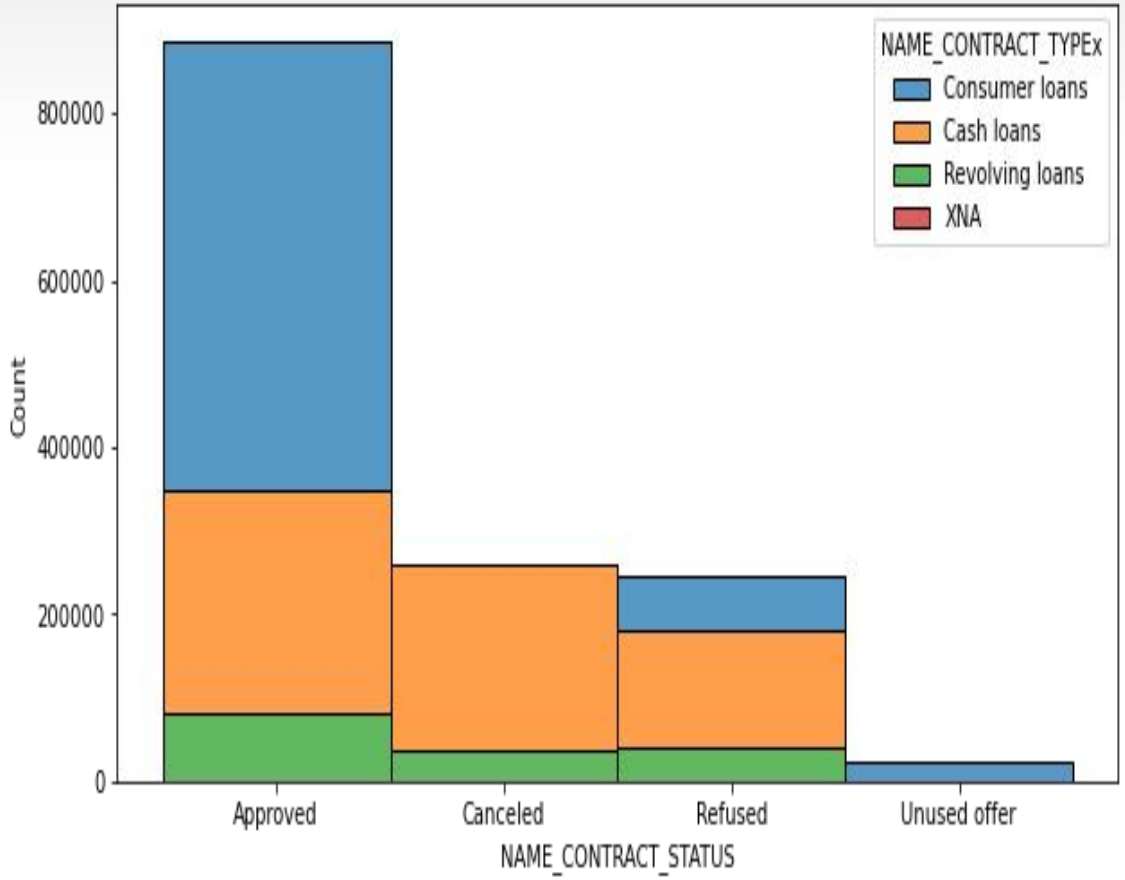
Laborers and Managers in difficulty are in in credit range of 0- 2000000 while Core Staff and High skill tech staff in difficulty are in credit range of 0-1500000.

Analysis After Merginging Of Current And Previous Loans Application.

Contract Status & Contract Type

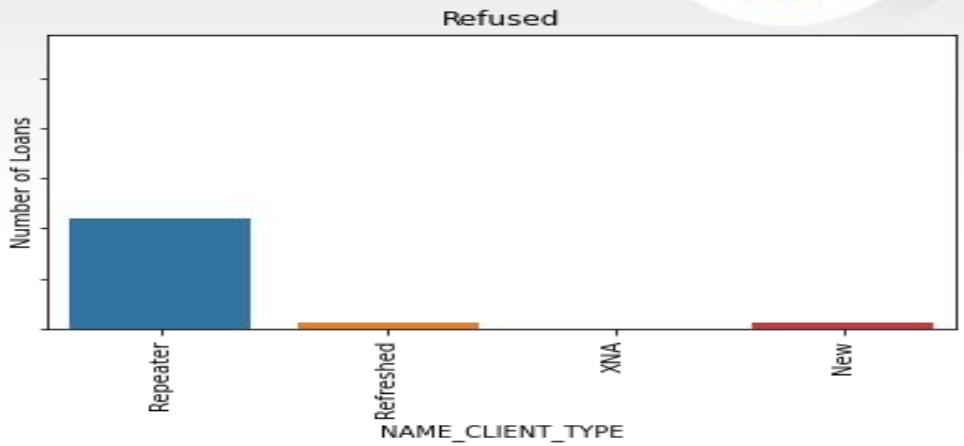


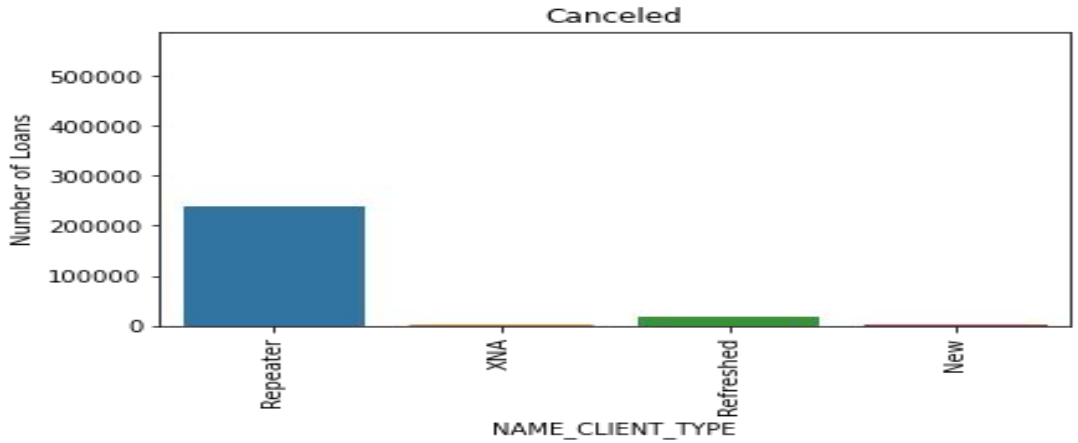
- Approved 62 %
- Canceled 18 %
- Refused 17 %
- Unused offer 2 %

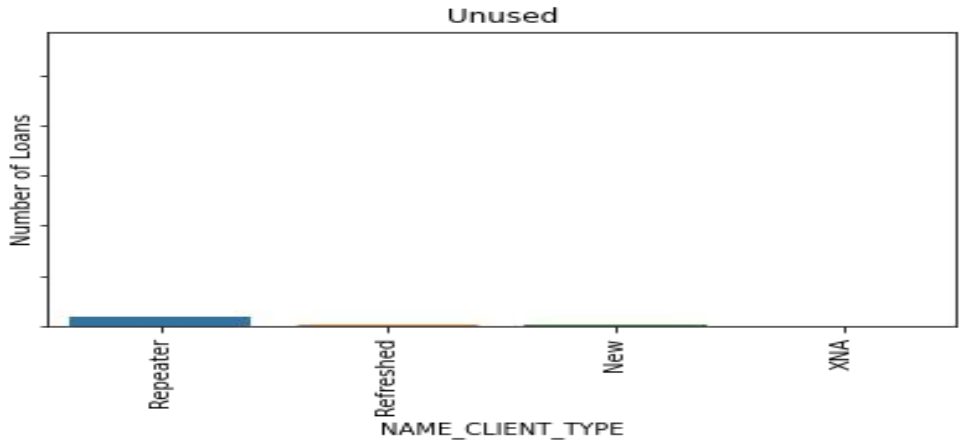


1. Here we see that most of the approved loans are Consumer loans and Cash loans.
2. Whole the most canceled loans are Cash loans.
3. The most refused loans are also Cash loans.

Client Type





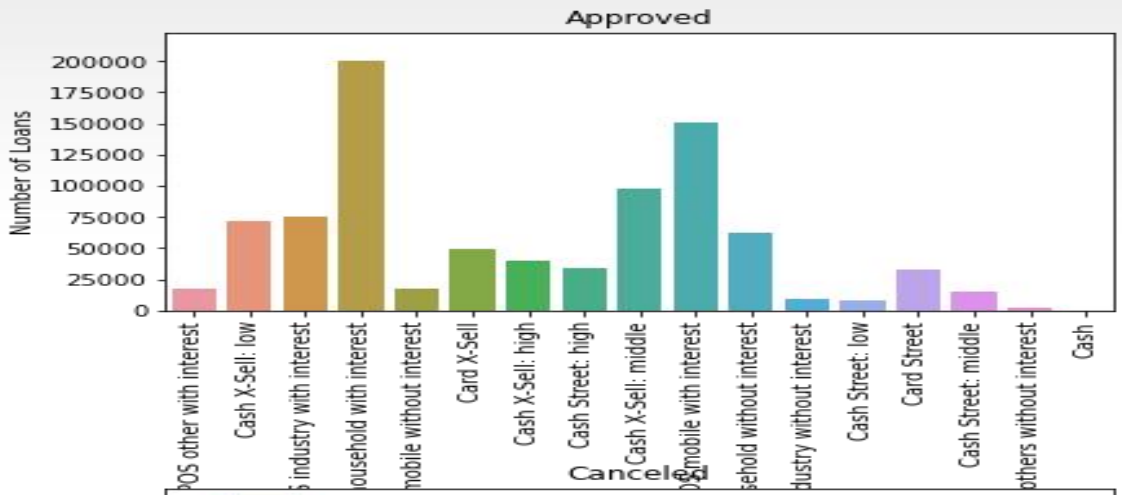


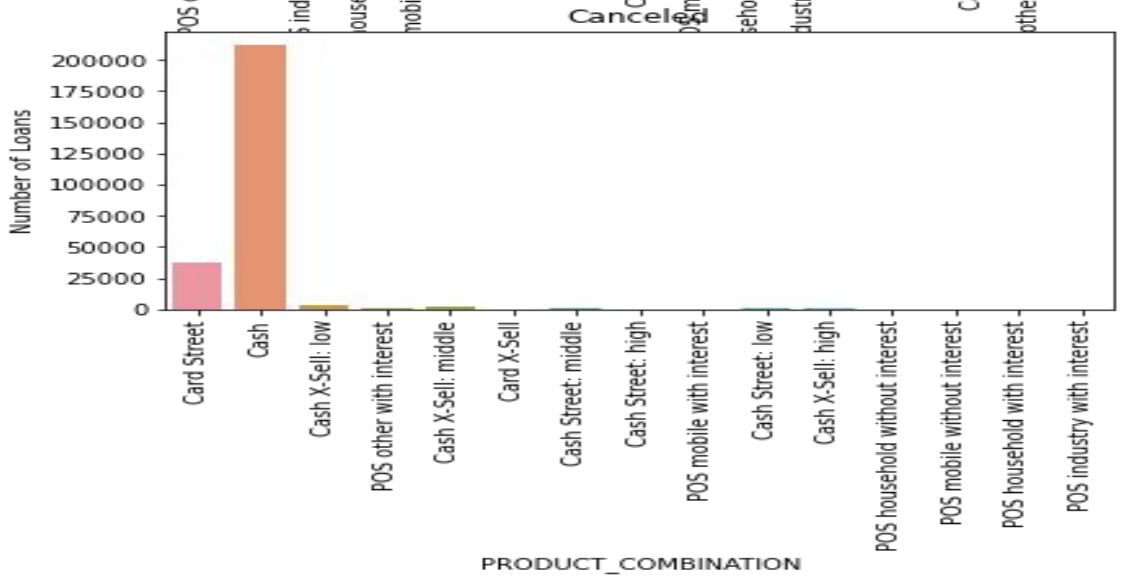
We can see that most approved loans are Repeater and New while most Refused and Canceled loans are repeater. New and refreshed loans are less refused.

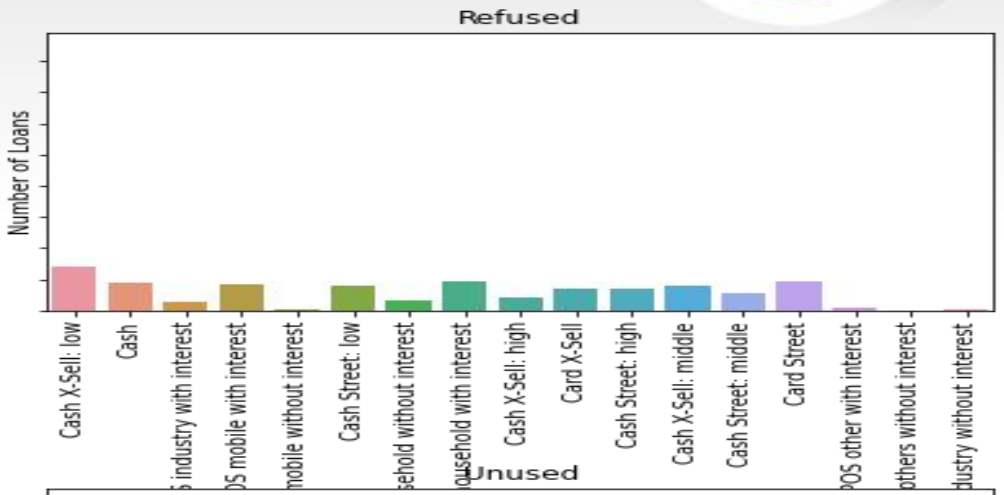
Total Percentage of Client Type

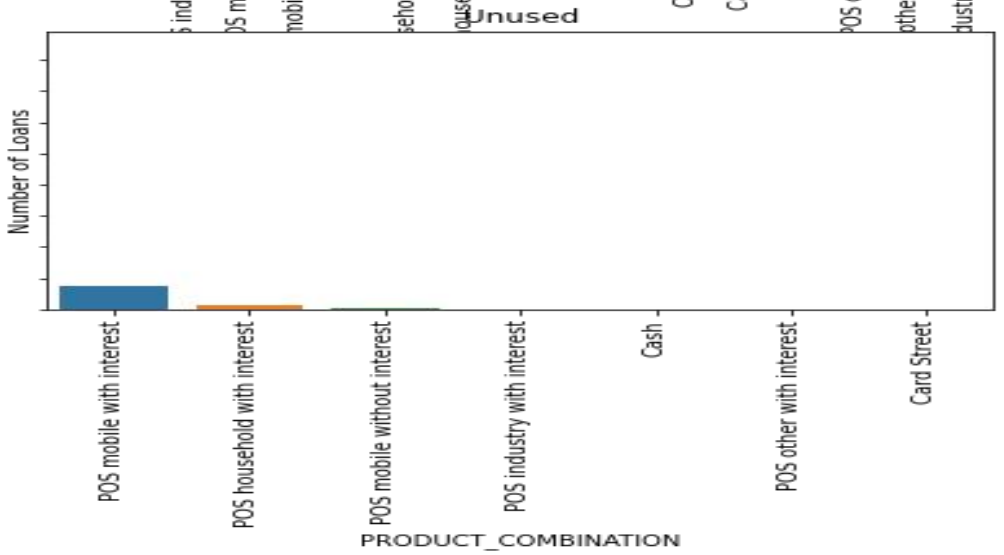
Repeater	73 %	New	18 %	Refreshed	8 %	XNA	0.2 %

Product Combination









The most approved loan here is POS household with interest. The Refused are cash x-sell:low and card street. And the most Canceled loan is Cash.

Result

- Bank should focus more on Commercial associate, Pensioner and State servant since overall there are more paying than defaulting.
- Businessman & Student do no to show defaults while maternity leave is less. Pensioners & State servant with default are in Income range of 500000. Unemployed with default are below 200000. Working People are the same in both category. Banks should be carefull with application in these catogories with this Income Range.
- Bank should focus more on Consumer loans and Second on Cash loans and lastly Revolving make a very small part of the loans.
- Bank should focus more on Females since they are a large part of the customer base. The female portion with difficulty in paying are less than those who have no difficulty in paying. While,it is the opposite for the males.
- People with House/apartments and working people are the most with succesfull payments and difficulties.
- Bank should focus more on people with Academic Degree & Higher education as there are more succefull payments than difficulties. It should focus less on Lower secondary as there is more payment difficulties.
- The customers with defaults are less in Academic degree at 500000 to 1500000. Higher Education people with defaults are around 2800000.

Result

- Bank should be carefull of single people and civil marriage since there are more unsuccesfull payment.
- In Occupation, Core staff, Managers and Accountants are less in difficulty are more are able to pay loans.
- Laborers and Managers with defaults are in in credit range of 0- 2000000 while Core Staff and High skill tech staff with defaults are in credit range of 0-1500000.
- People above 40 years of employment pay all thier installments.
- Bank should focus more on older people as they are able to Pay installments more than younger people.
- Most of the defaulters are people with total Income below 500000 & Credit below 1500000 Loan Amounts.
- From the Previous Application, We learned that Banks should focus more on Consumer Loans since Most of the Refused and Canceled loans are Cash loans.
- The most approved loan is POS household with interest. The Refused are cash x-sell:low and card street. And the most Canceled loan is Cash.