



Credit EDA Case Study

BY – SOURAB KADADI

Introduction

- ▶ The Case Study aims to give an idea of applying EDA in a real business scenario.

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 as 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.
- ▶ 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.

Business Understanding

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

- ▶ **Approved:** The Company has approved loan Application
- ▶ **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.
- ▶ **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
- ▶ **Unused offer:** Loan has been cancelled by the client but on different stages of the process.

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 utilise 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).

Database considered for Analysis

- ▶ 1. '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**.
- ▶ 2. '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**.
- ▶ 3. 'columns_description.csv' is data dictionary which describes the meaning of the variables.

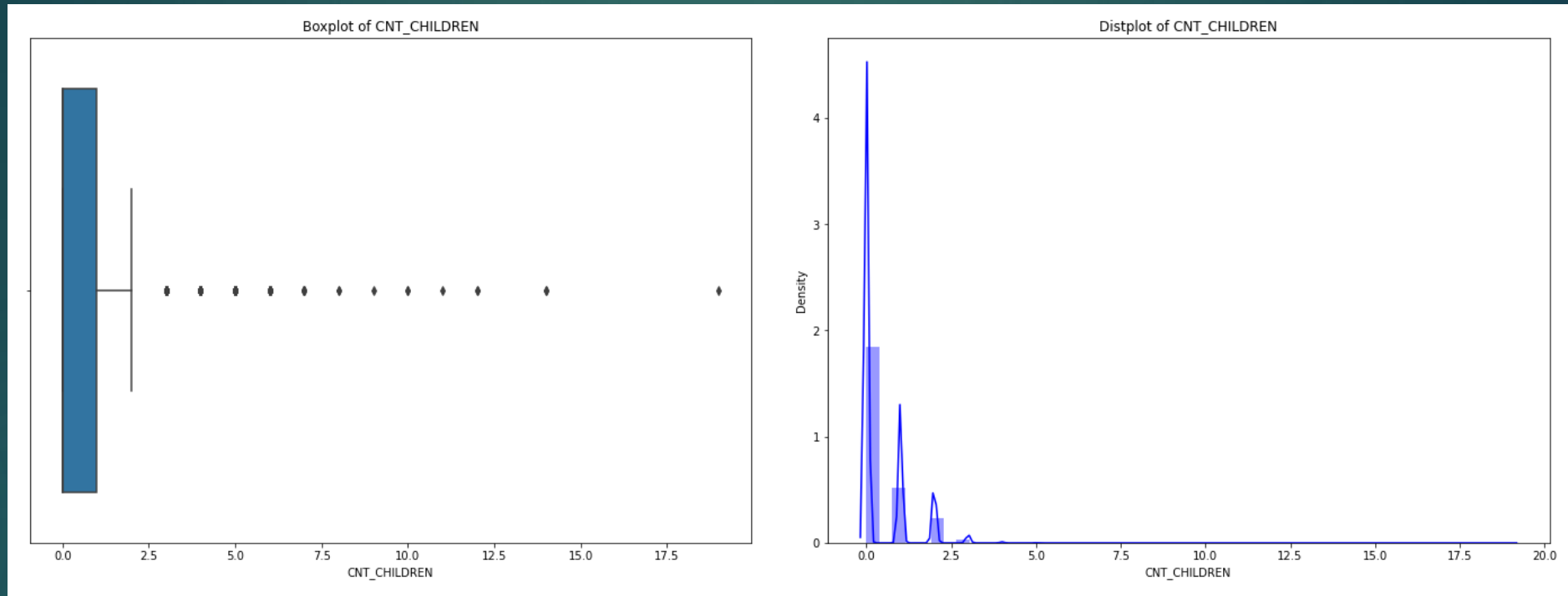


Analysis of the Information from Application Database



Outlier Analysis

Analysis of `CNT_CHILDREN`



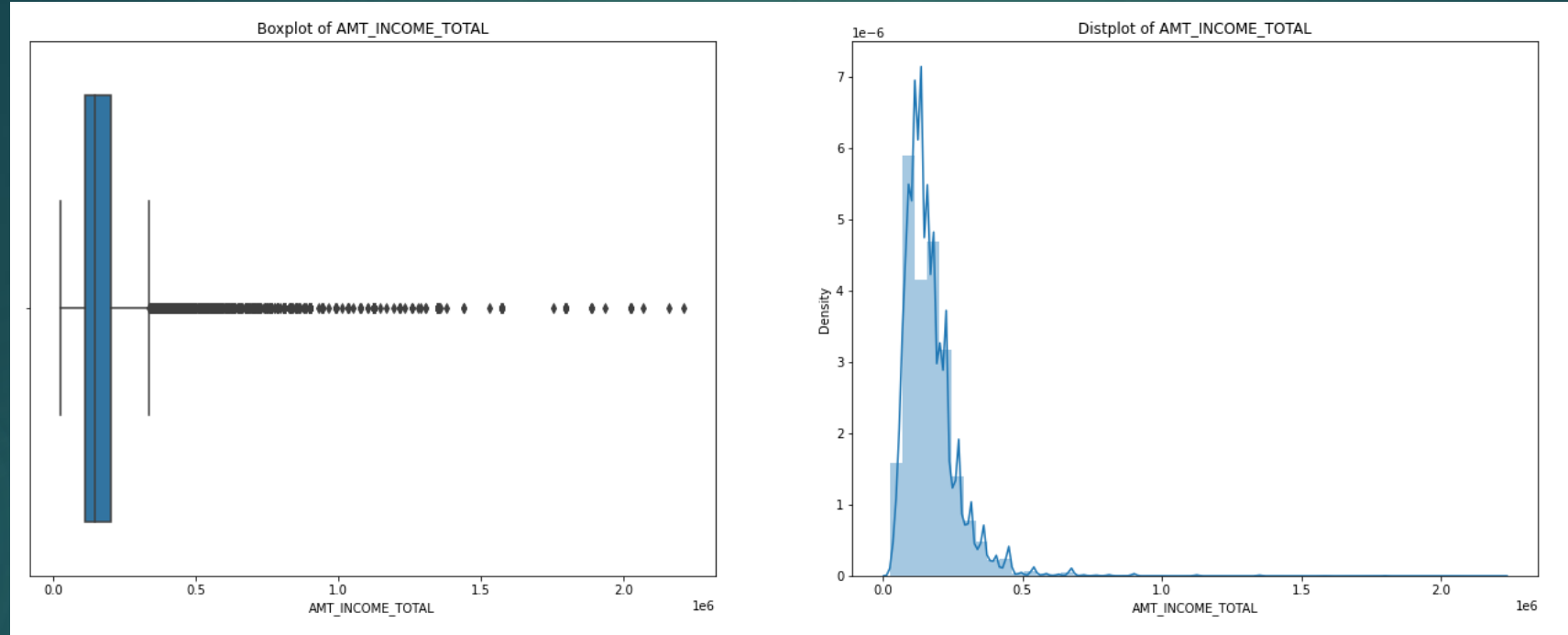
Observations

1. From the above, Applicants with children above 2.5 are outliers and are very minimal

Conclusion

1. Applicants with more than 3 children are outlier cases

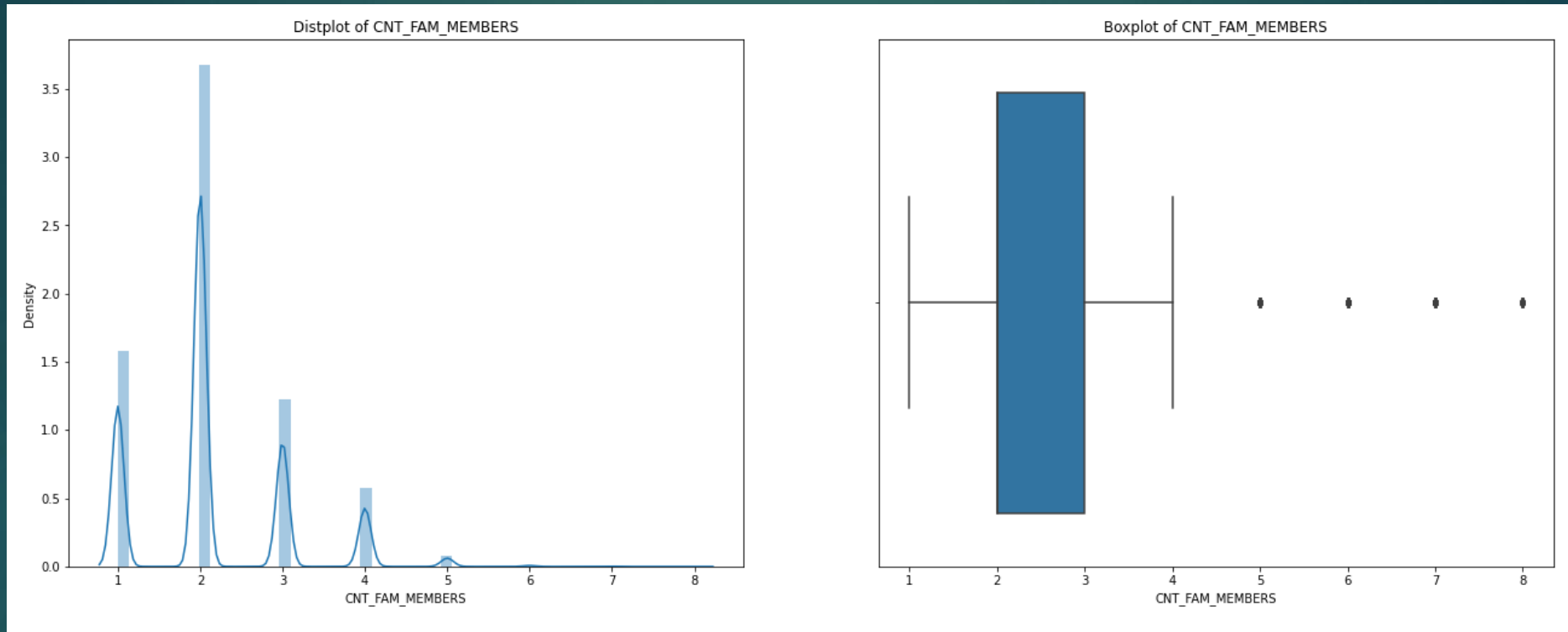
Analysis of `AMT_INCOME_TOTAL`



Observations

Applicants with income of 2250000 are clearly outliers and are at 99.99 percentile

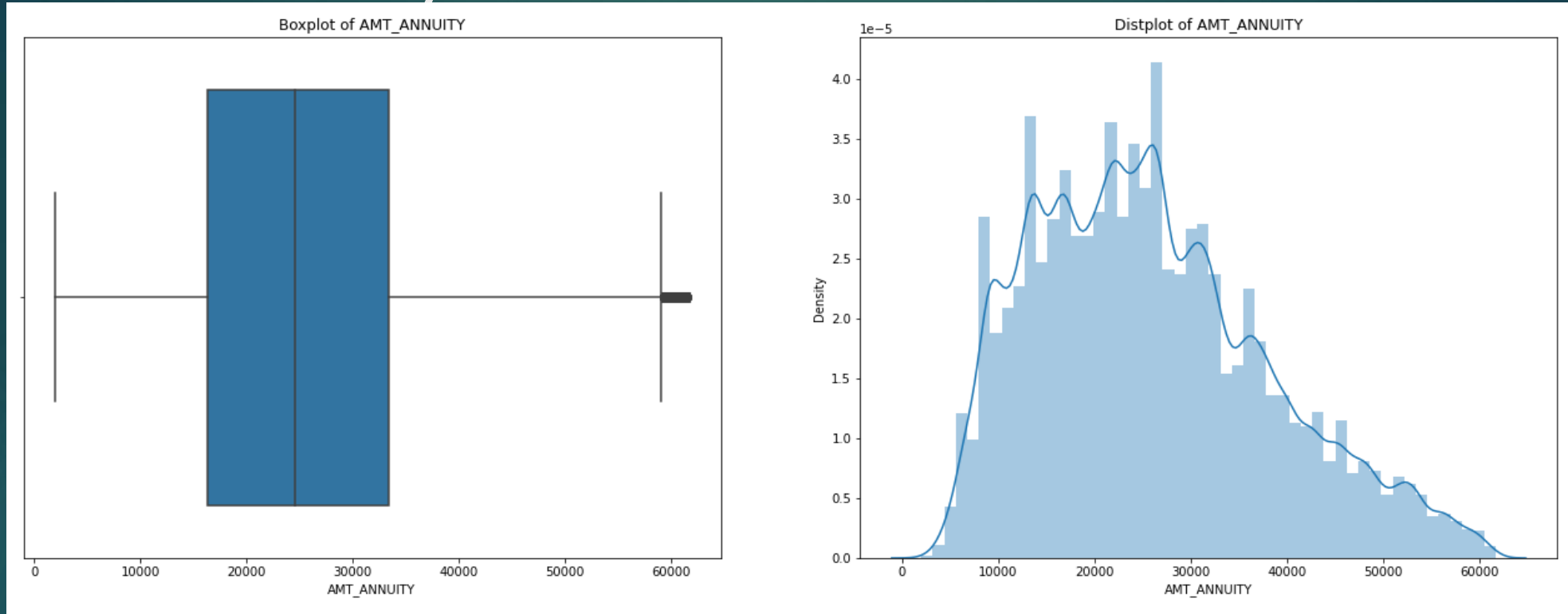
Analysis of `CNT_FAM_MEMBERS`



Observations

Applicants with more than 5 family members are outliers

Analysis of `AMT_ANNUITY`



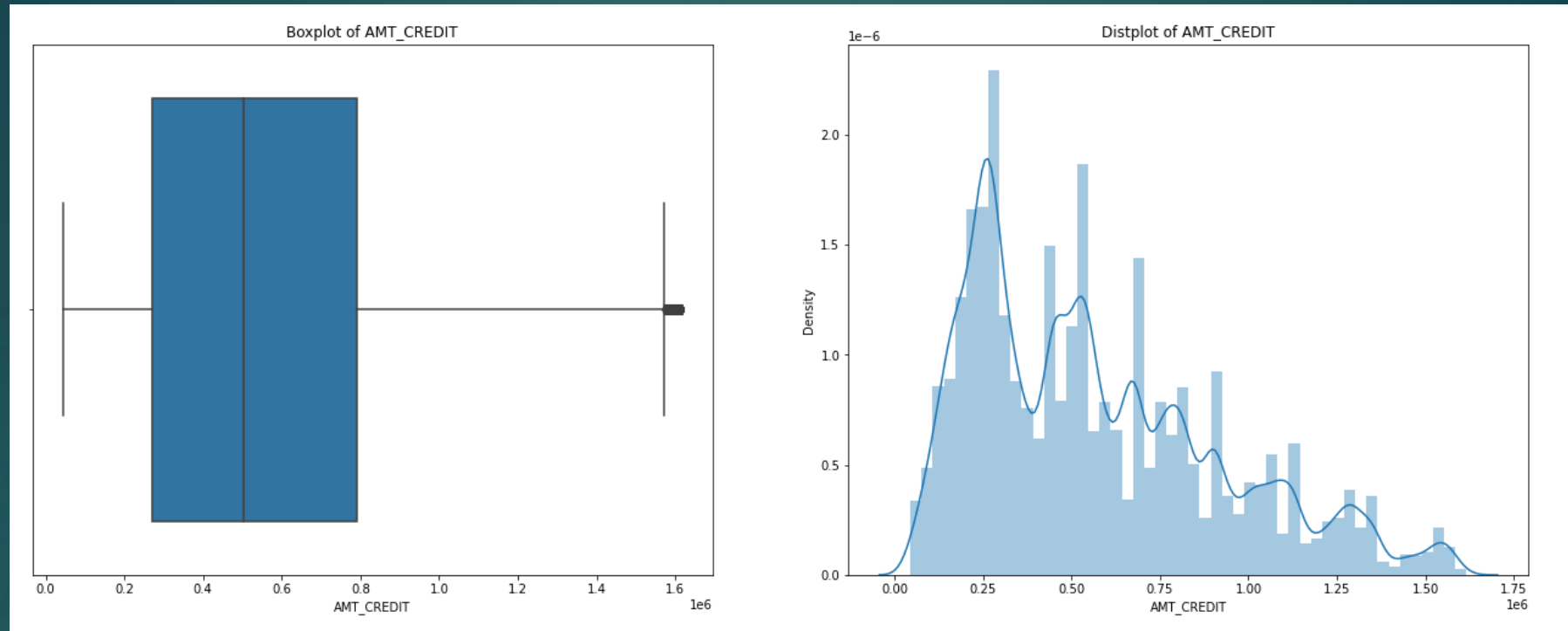
Observations

The outliers exist after 61715.25 (Outlier value is derived using Max_value using IQR formula)

Conclusion

Applicants with AMT_ANNUITY above 61715.25 (calculated using IQR) are outliers

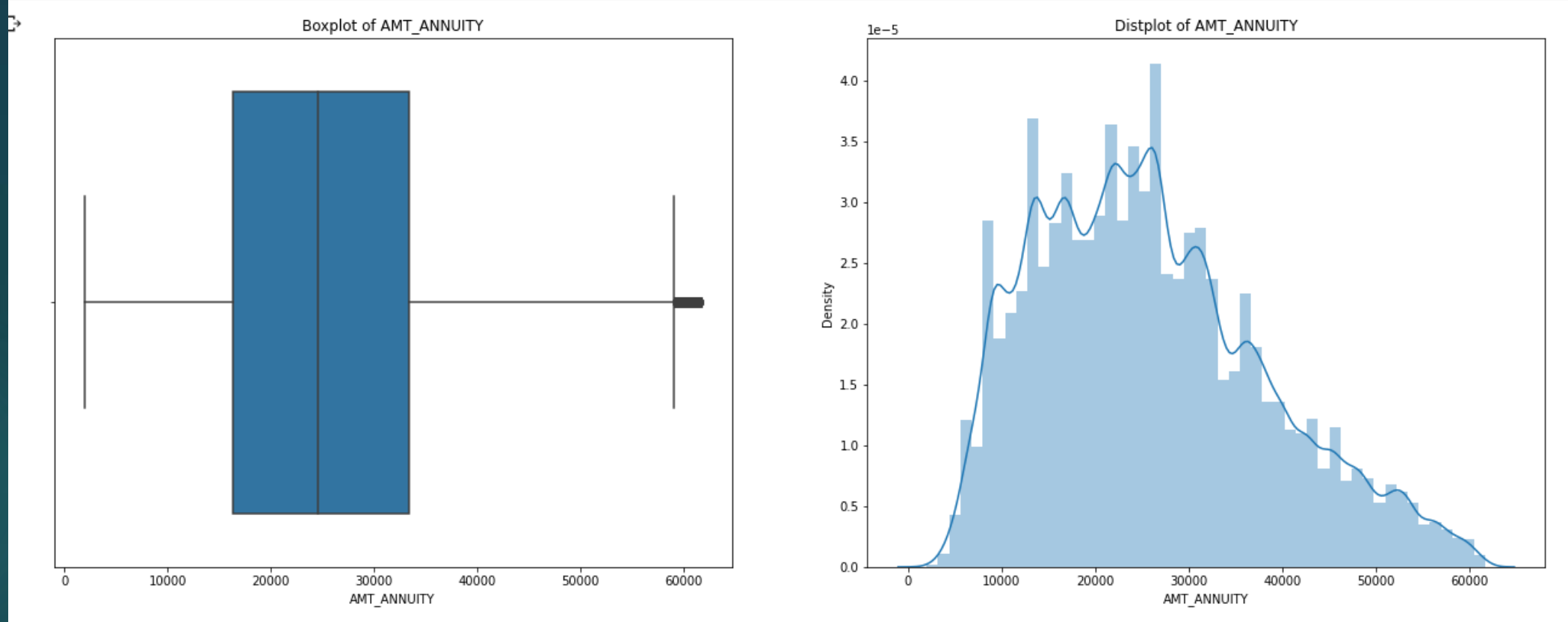
Analysis of `AMT_CREDIT`



Observations

From the above it is clear that credit amount above 1616625.0 derived using the `MAX_VALUE` in IQR is a outlier

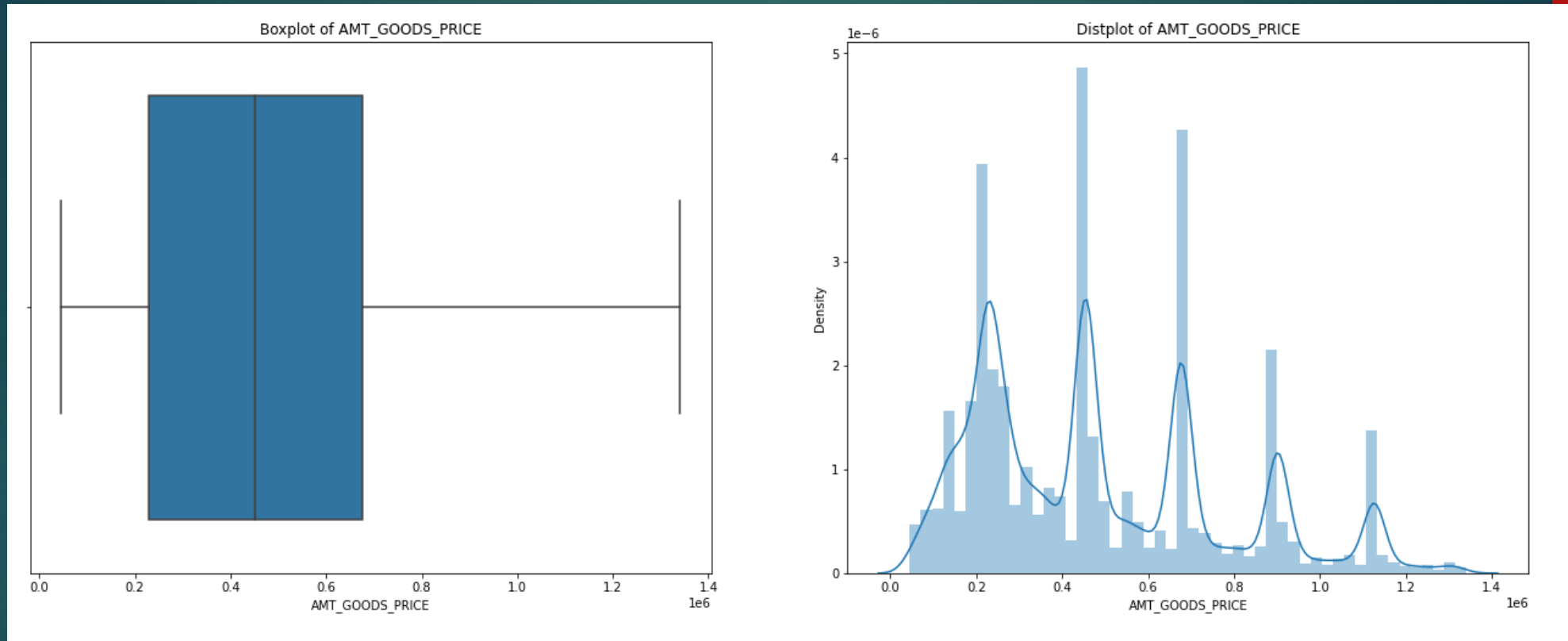
Analysis of `AMT_ANNUITY`



Observations

From the above plots it is clear that the outlier exists after 61715.25 which is also derived from `MAX_VALUE` in IQR

Analysis of `AMT_GOODS_PRICE`



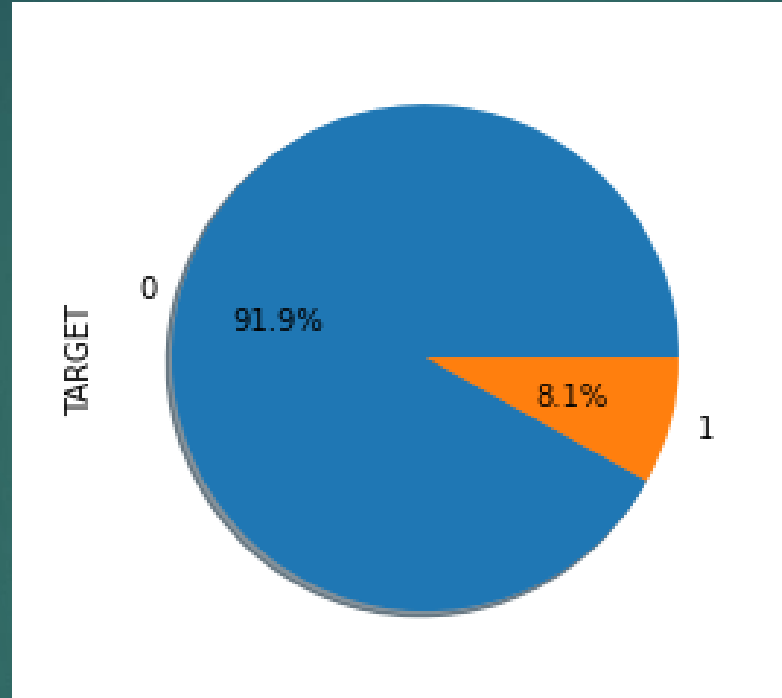
Observations

From the above plots it is clear that the outlier exists after 1341000.0 which is also derived from `MAX_VALUE` in IQR



Checking for the Imbalance of The **Target** Variable

Analysis of `TARGET`



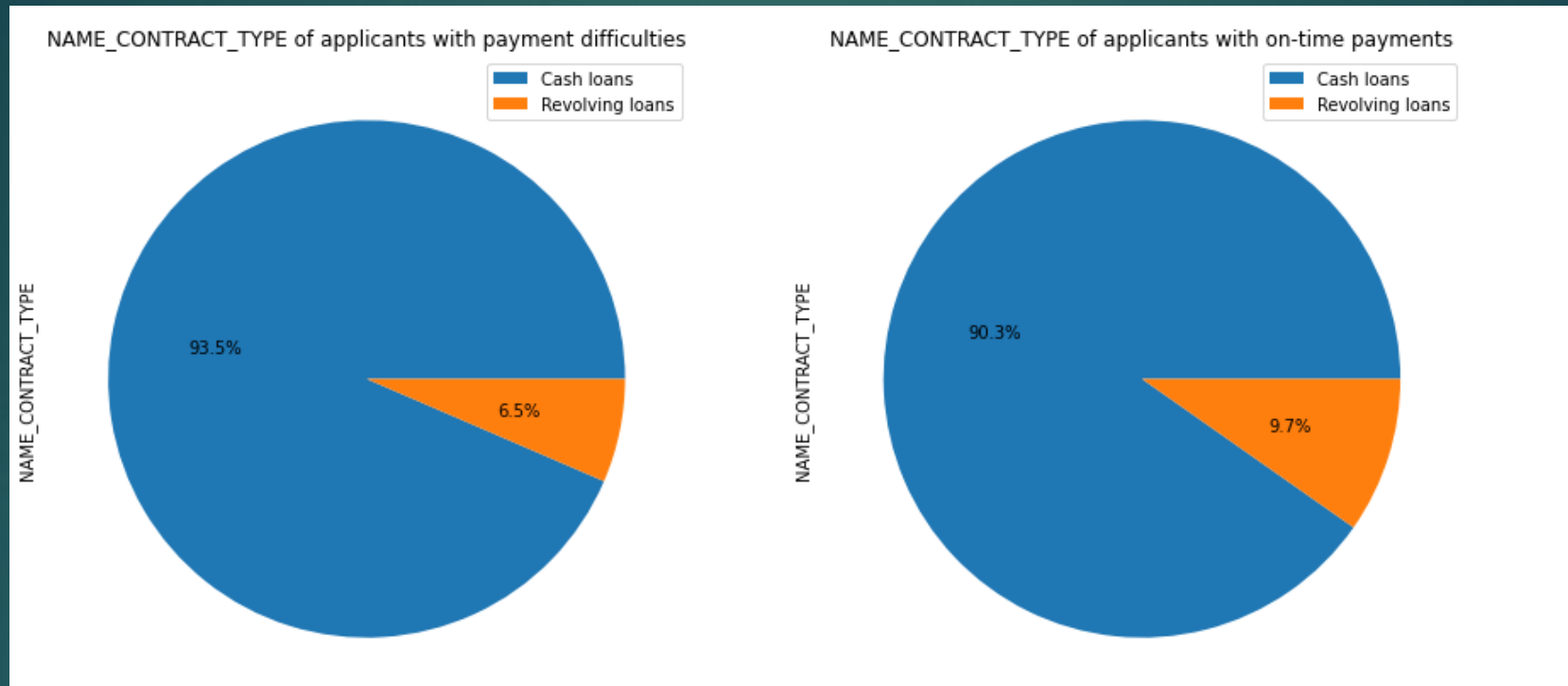
Observations

- There is imbalance in `TARGET` variable based on the % of observations
- `TARGET` value 1 represents client with payment difficulties (he/she had late payment more than X days on at least one of the first Y installments of the loan) which is 8.1% of the data
- `TARGET` value 0 represents all other cases than 1. This is 91.9% of the data



Univariate Analysis of Categorical Variables

Analysis of `NAME_CONTRACT_TYPE`



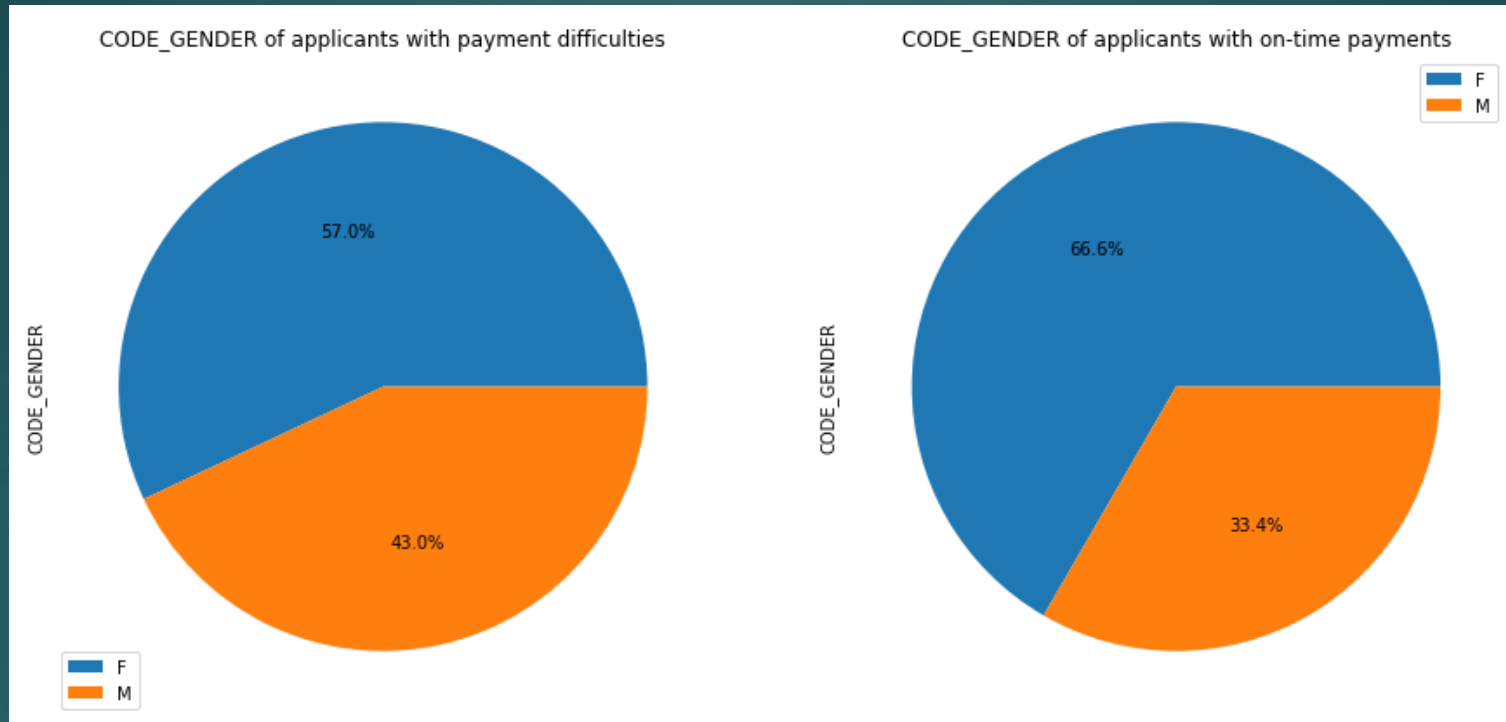
Observations

There is no significant differences in `NAME_CONTRACT_TYPE` b/w applicants with payment difficulties and on-time payments for both the CountPlot and Piechart

Conclusion

`NAME_CONTRACT_TYPE` column does not provide any conclusive evidence in favor of applicants with payment difficulties or on-time payments

Analysis of `CODE_GENDER`



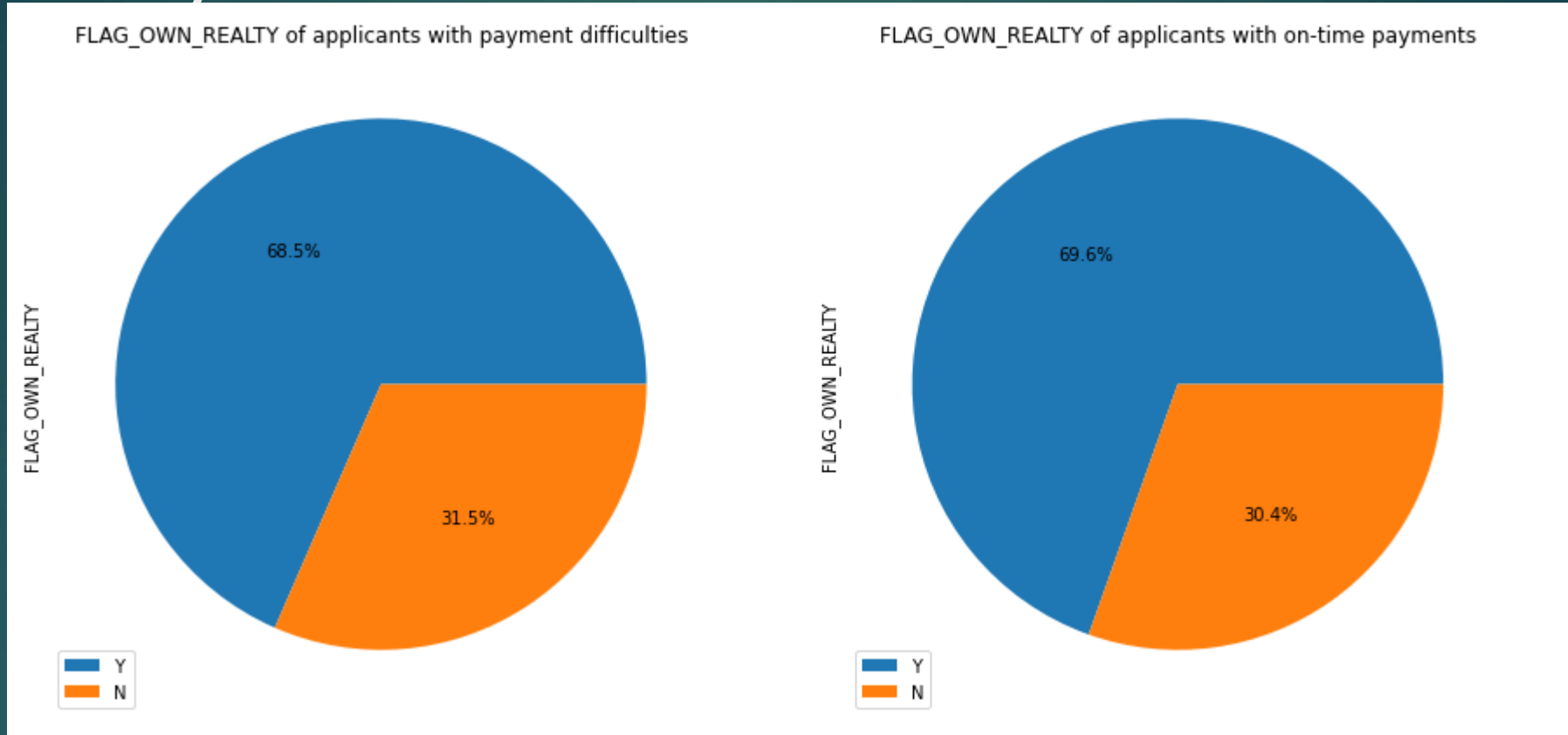
Observations

- It can be observed from the pie chart that, the percentage of on time payments of male increases by around 9% from on-time payments to applicants with payment difficulties for male.

Conclusion

- There is a weak inference from the `CODE_GENDER` column that males have higher payment difficulty than Female

Analysis of `FLAG_OWN_REALTY`



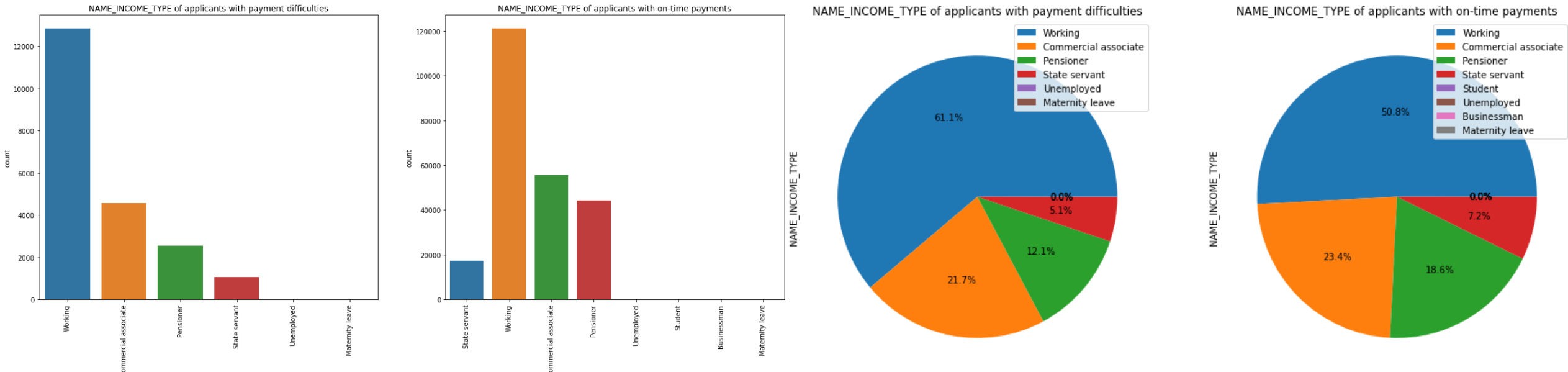
Observations

There is no significant differences in `FLAG_OWN_REALTY` b/w applicants with payment difficulties and on-time payments for both the CountPlot and Piechart

Conclusion

`FLAG_OWN_REALTY` column does not provide any conclusive evidence in favor of applicants with payment difficulties or on-time payments

Analysis of `NAME_INCOME_TYPE`



Observations

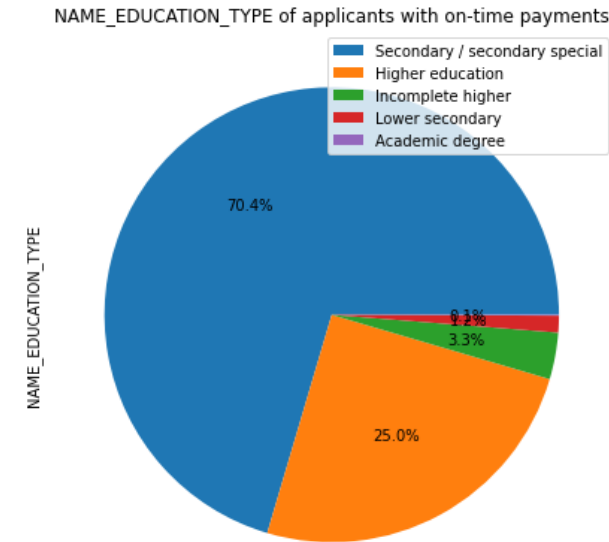
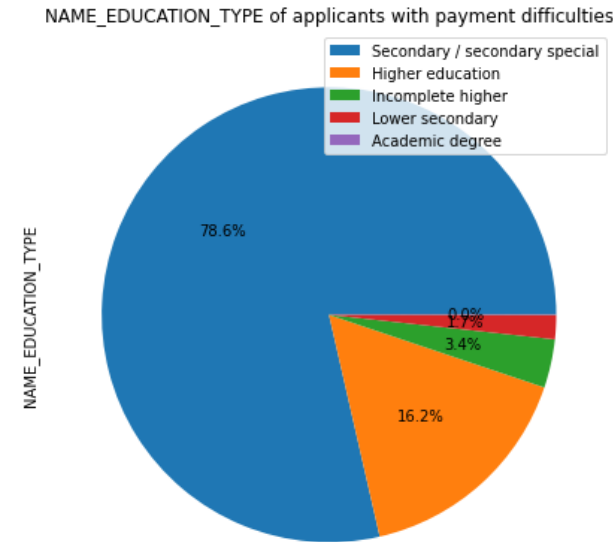
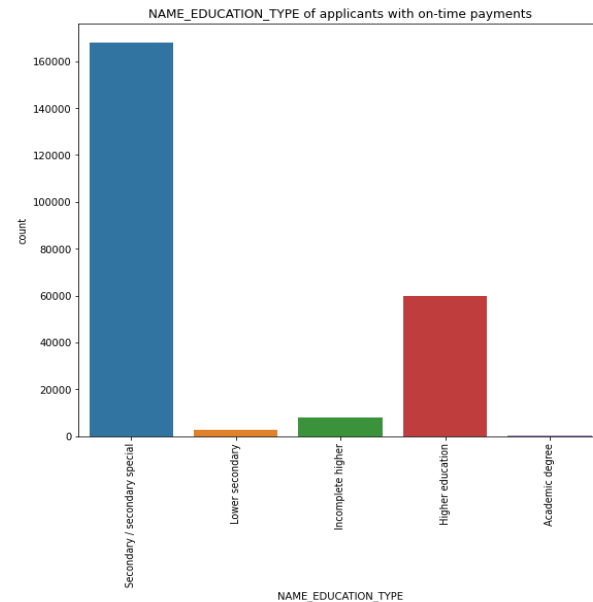
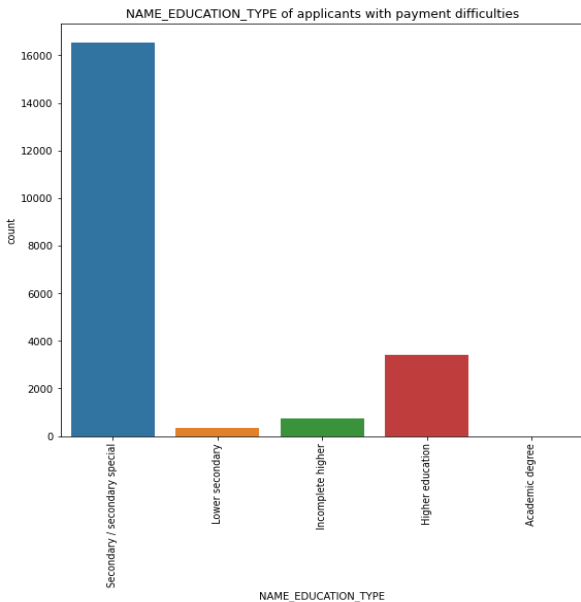
From the Piechart for `NAME_INCOME_TYPE`, we can observe that, working applicants have higher payment difficulties in comparison to commercial associate, pensioner and state servant.

The working applicants were 50.8% of the total who had made ontime payments whereas the same increased to 61.1% for the payment with difficulties.

Conclusion

we can conclude that the working applicants have payment with difficulties.

Analysis of `NAME_EDUCATION_TYPE`



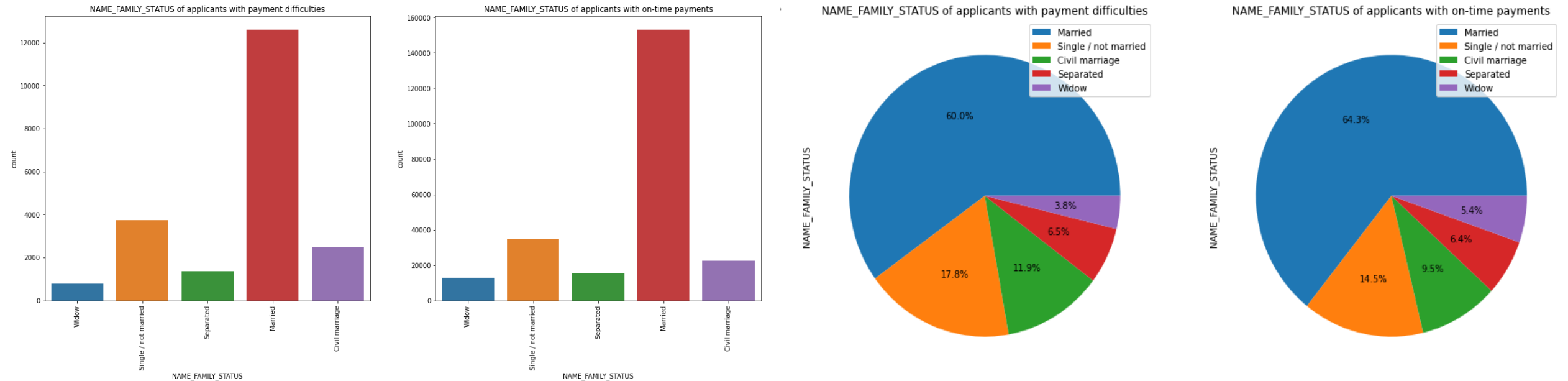
Observations

For `NAME_EDUCATION_TYPE` column, the applicants with Secondary Education level have a slightly higher chances of having payment with difficulties as indicated in the pie chart with increase in contribution to the pool of total applicants of Secondary Education from 70.4% in on-time payments to 78.6% to payment with difficulties.

Conclusion

For `NAME_EDUCATION_TYPE` column, the applicants with Secondary Education level have weak inference for payment with difficulties.

Analysis of `NAME_FAMILY_STATUS`



Observations

- Applicants who are 'Married' are 64.3% with on-time payments and 60.0% with payment difficulties
- Applicants who are 'Single/Not Married' are 14.5% with on-time payments and 17.8% with payment difficulties

Conclusion

- Applicants who are 'Married' or 'Widow' make on-time payments better comparatively. However, this is a weak correlation.
- Applicants who are 'Single/not married' have more difficulties with on-time payments comparatively. However, this is a weak correlation.



Co-relation for
numerical columns for
both the cases
0 and 1 of TARGET variable

Co-relation for applicants

Applicants with On time payments

OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	1.00
AMT_CREDIT	AMT_GOODS_PRICE	0.99
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.95
CNT_CHILDREN	CNT_FAM_MEMBERS	0.88
...		
REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	-0.54
FLAG_EMP_PHONE	FLAG_DOCUMENT_6	-0.60
	DAYS_BIRTH	-0.62
DAYS_EMPLOYED	FLAG_EMP_PHONE	-1.00

Applicants with Payment difficulties

OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	1.00
AMT_CREDIT	AMT_GOODS_PRICE	0.99
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.95
CNT_CHILDREN	CNT_FAM_MEMBERS	0.88
...		
REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	-0.54
FLAG_EMP_PHONE	FLAG_DOCUMENT_6	-0.60
	DAYS_BIRTH	-0.62
DAYS_EMPLOYED	FLAG_EMP_PHONE	-1.00

Observations

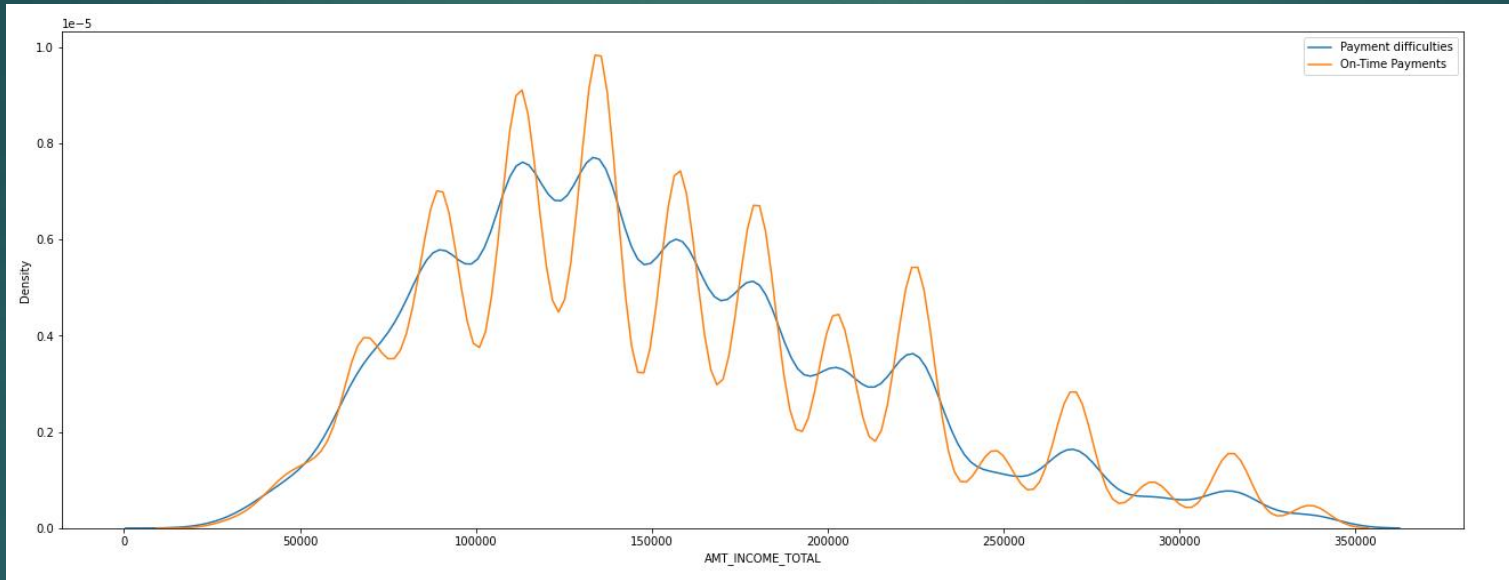
The Co-relation top 5 +ve and top 5-ve are same for applicants with on-time payments and applicants with payment difficulties



Univariate analysis of Numerical variables

Analysis of AMT_INCOME_TOTAL

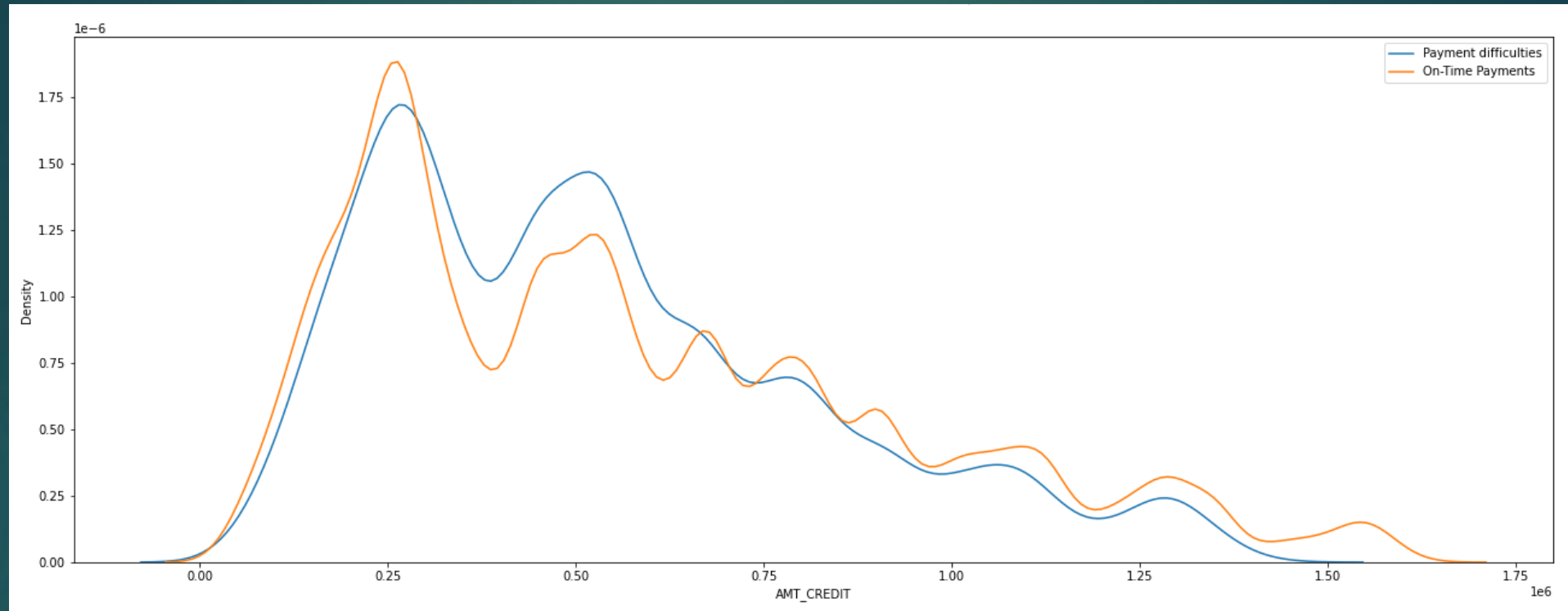
Identifying outlier in both the cases of TARGET variable for AMT_INCOME_TOTAL



Observations

The TARGET column with payment difficulties has a normal distribution whereas the On-time payments display erratic spikes which has to be analysed further.

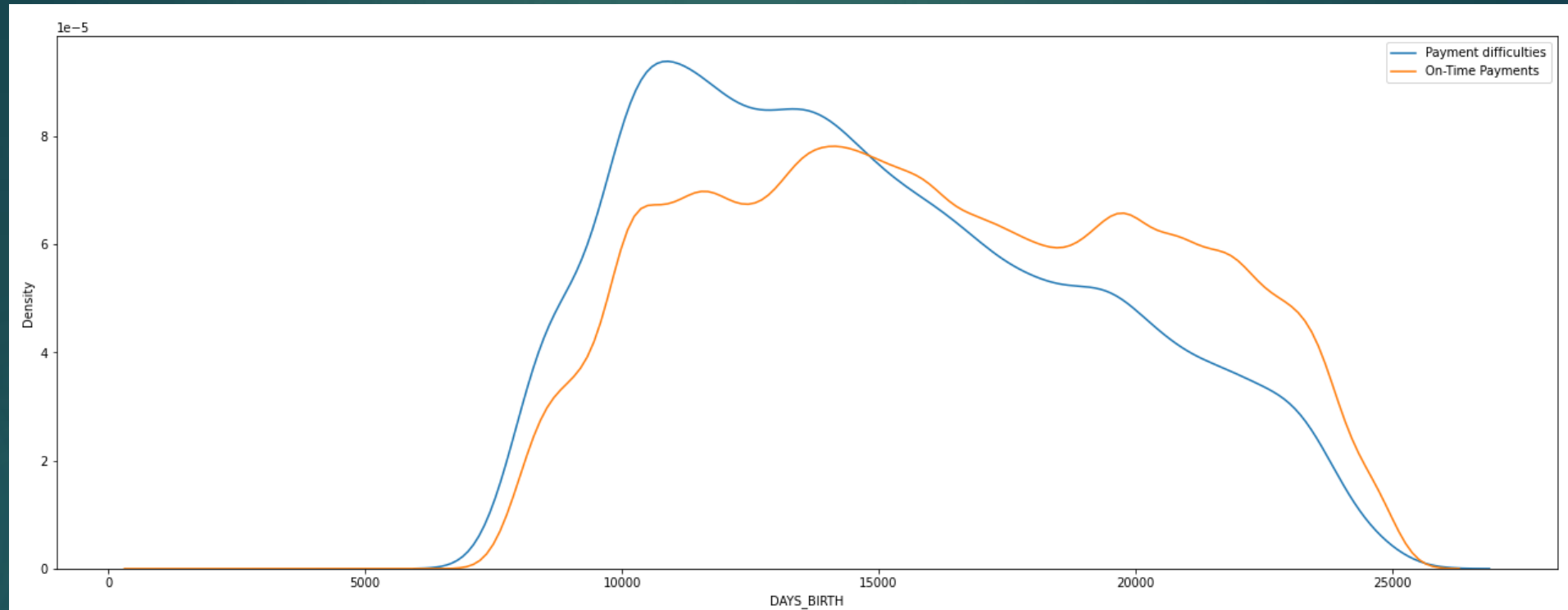
Analysis of `AMT_CREDIT`



Observations

- For `AMT_CREDIT` between 250000 and approximately 650000, there are more applicants with Payment difficulties
- For `AMT_CREDIT` > 750000 and `AMT_CREDIT` < 250000, there are more applicants with On-Time Payments

Analysis of `DAYS_BIRTH`

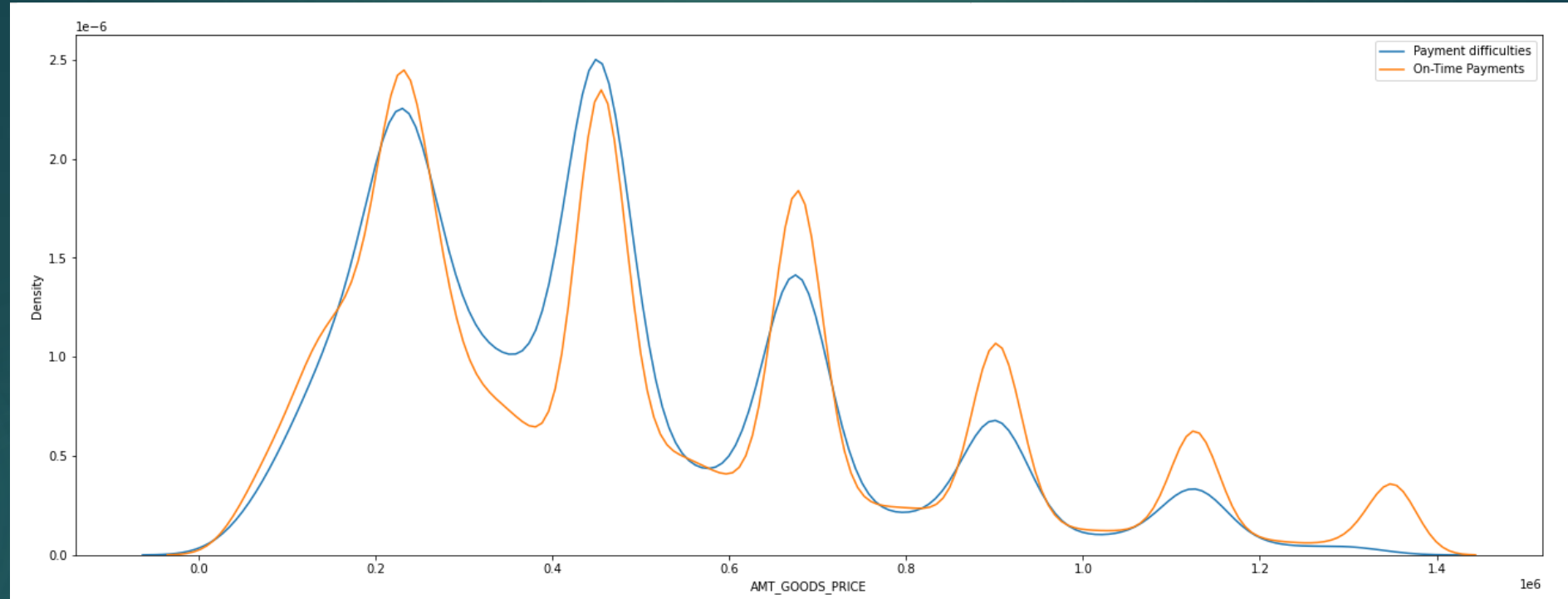


Observations

For `DAYS_BIRTH` between 6500 days and 15000 days which is equivalent to 17 years to 41 years, there are more applicants with Payment difficulties

On the other way, for `DAYS_BIRTH` > 15000 days which is around 41.1 years, there are more applicants with On-Time Payments

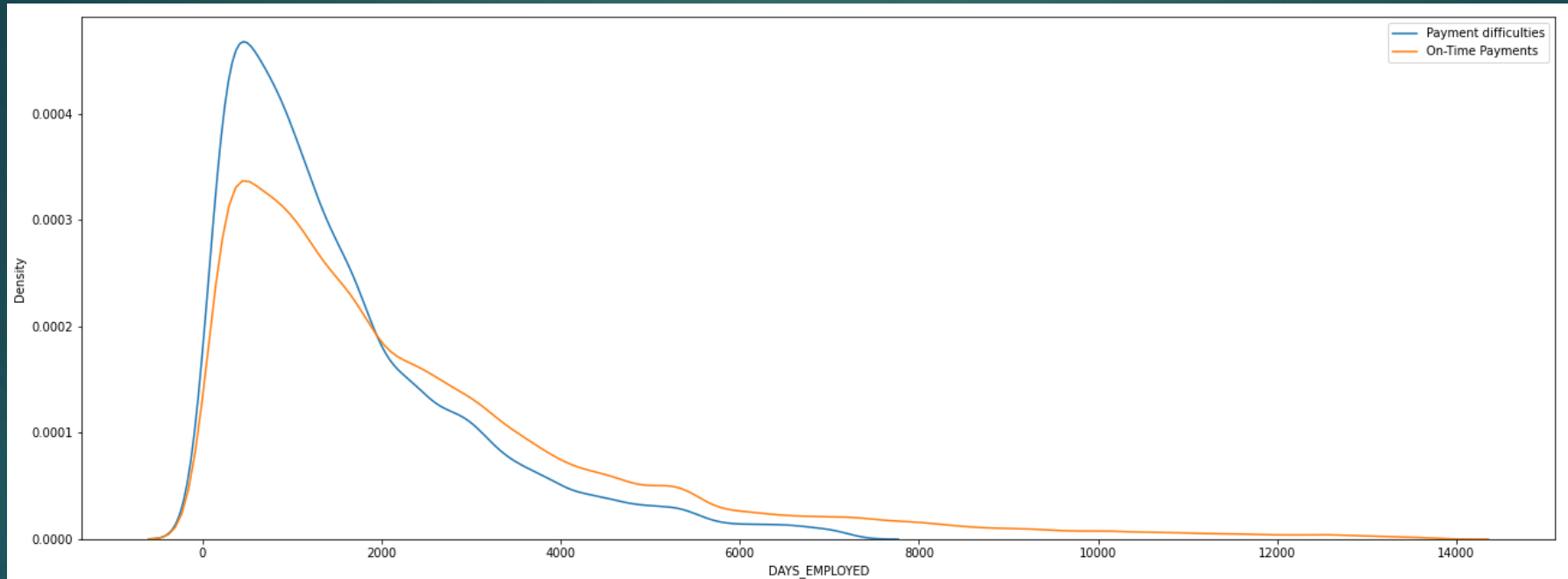
Analysis of `AMT_GOODS_PRICE`



Observations

- For `AMT_GOODS_PRICE` between 200000 and 550000 there are more applicants with Payment difficulties
- On the other way, for `AMT_GOODS_PRICE` > 550000, there are more applicants with On-Time Payments

Analysis of `DAYS_EMPLOYED`



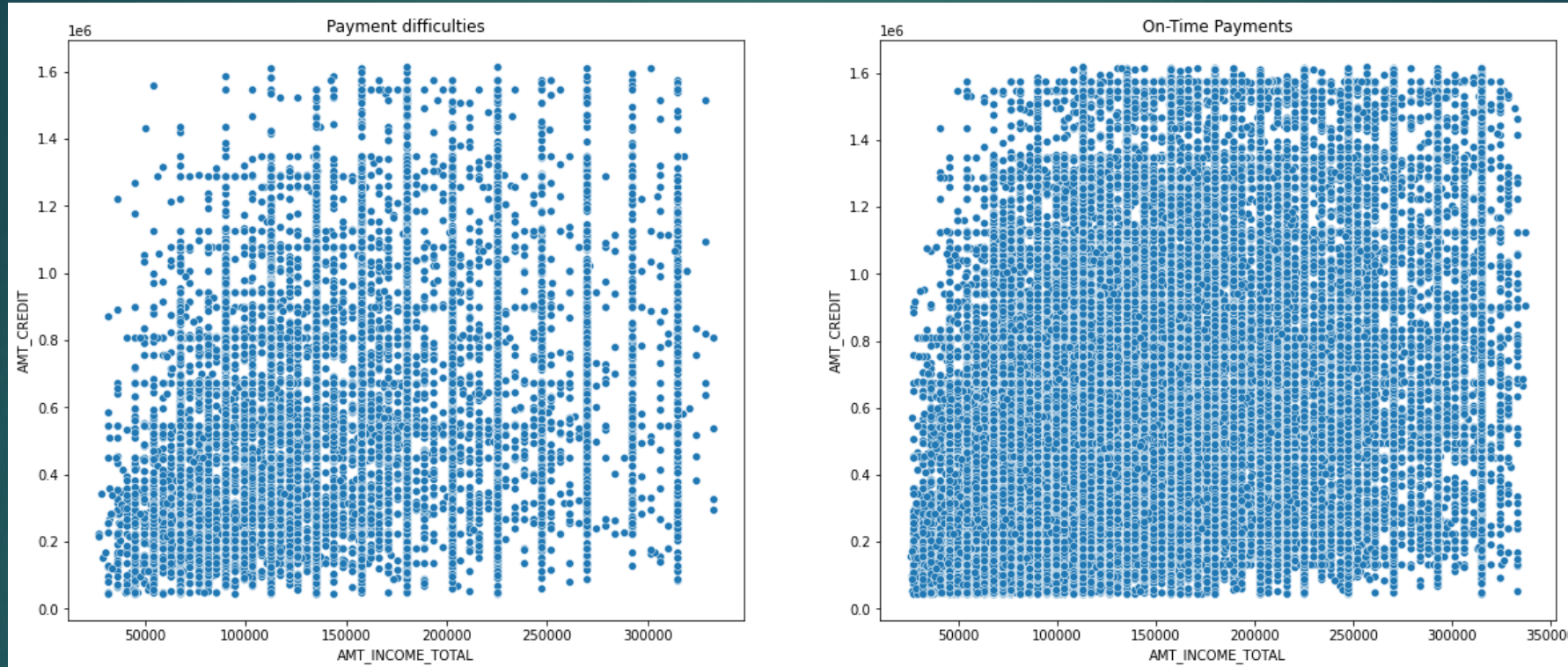
Observations

- For `DAYS_EMPLOYED` less than 2000 days which is equivalent to around 5.5 years, there are more applicants with Payment difficulties
- On the other way, for `DAYS_EMPLOYED` > 2000 days which is around 5.5 years, there are more applicants with On-Time Payments



Bi-Variate Analysis

Analysis of `AMT_INCOME_TOTAL` V/S `AMT_CREDIT`



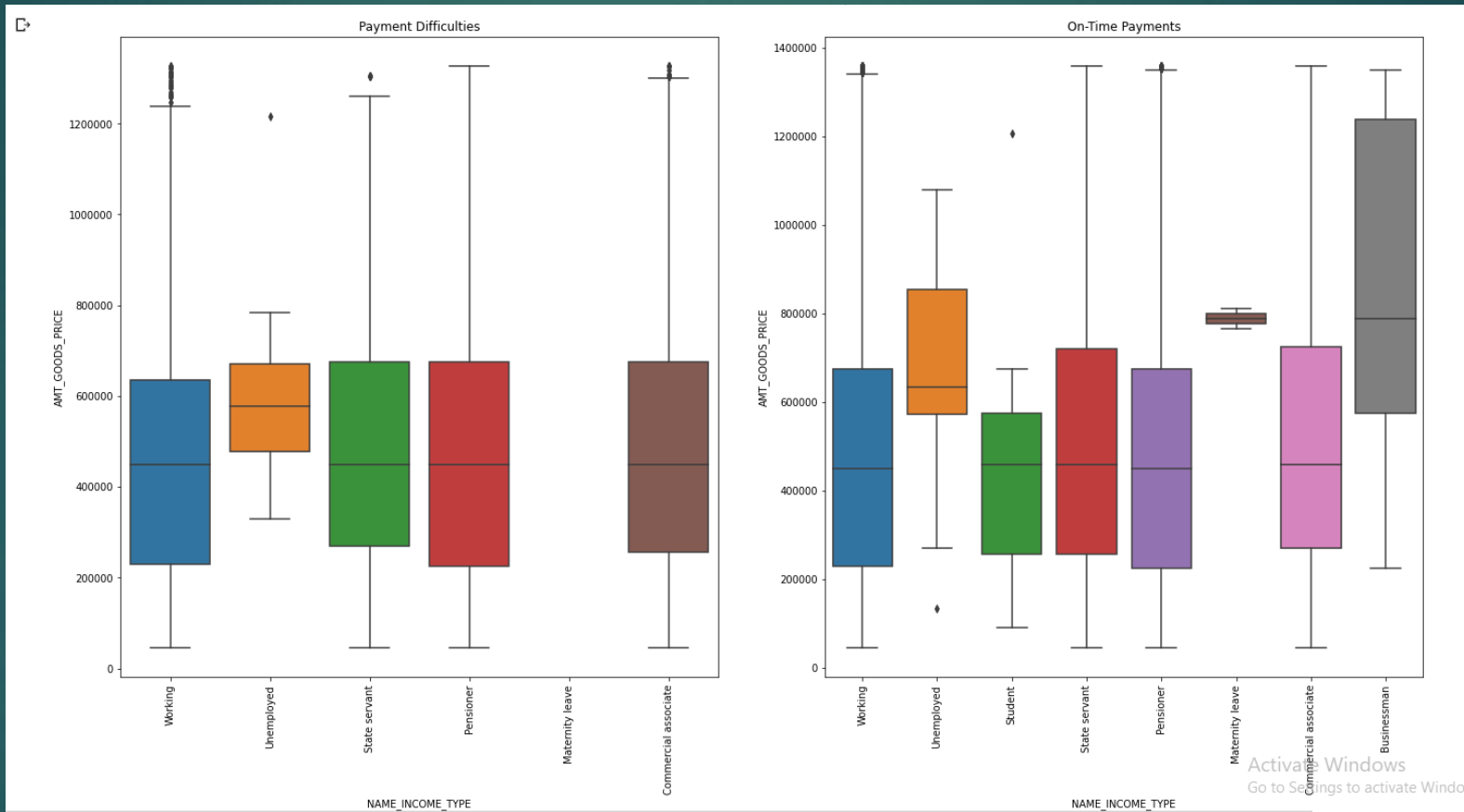
Observations

It is difficult to spot the trend for AMT_INCOME_TOTAL Vs. AMT_CREDIT



Continuous V/S Categorical variables

Analysis of NAME_INCOME_TYPE V/S AMT_GOODS_PRICE V/S CODE_ GENDER





Conclusion

Target categories for loan disbursal

Applicants who are above 22 years and below 50 years

Applicants who are married

Male applicants with academic background