❑ Problem Statement

- It is important to note that this data carries two risks

1. The Bank can suffer if the customer is unable to handle the load and default it
2. BANK will lose money if it does not give the loan to the customer who is able to repay it.

❑ Assumptions

- I am getting only one approximation in this given set, as I checked
- According to my observation, XNA is present in each of these columns (ORGANIZATION_TYPE, NAME_GOODS_CATEGORY, NAME_SELLER_INDUSTRY, NAME_CASH_LOAN_PURPOSE, NAME_PORTFOLIO, NAME_PAYMENT_TYPE, CODE_GENDER)

❑ Data Cleaning & Fill Missing Value

- First of all in EDA I cleaned the null values and get its percentage, and removed from that which were more than 35% NULL values, because if we fill that value, we may have problem in visualization.

- After that, I removed the low percentage NULL value (less than 1%) because we have a lot of data,

- I filled the mean value in place of NA, which were column integers, floating, and I filled the mode value in the categorical (object) column.

- I removed outliers with the help of boxplot

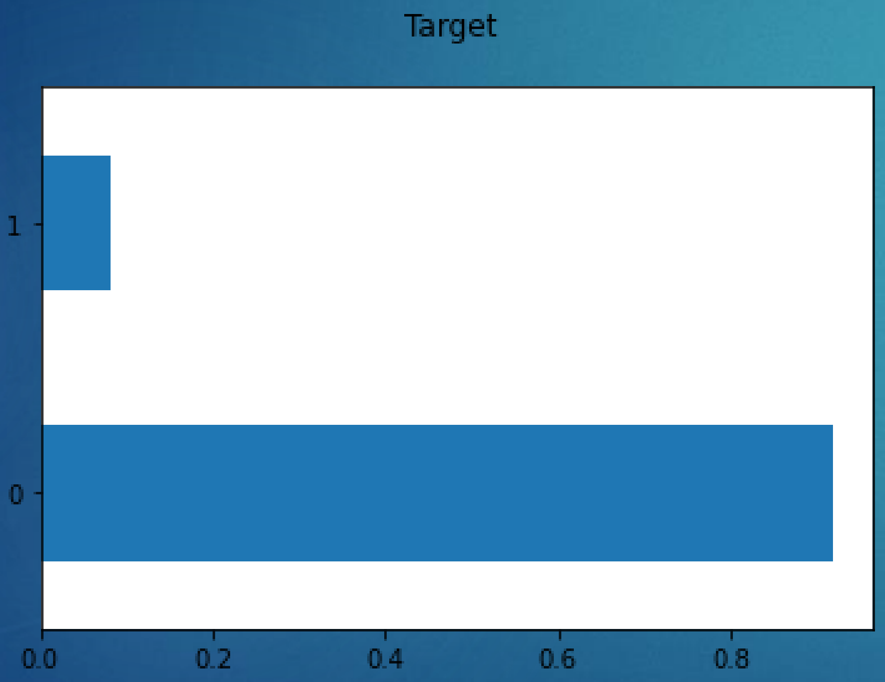
- In standardizing numerical values, I checked the datatype of the column, all the columns were correct, but many columns had negative values but they should have been positive, I converted it into positive.

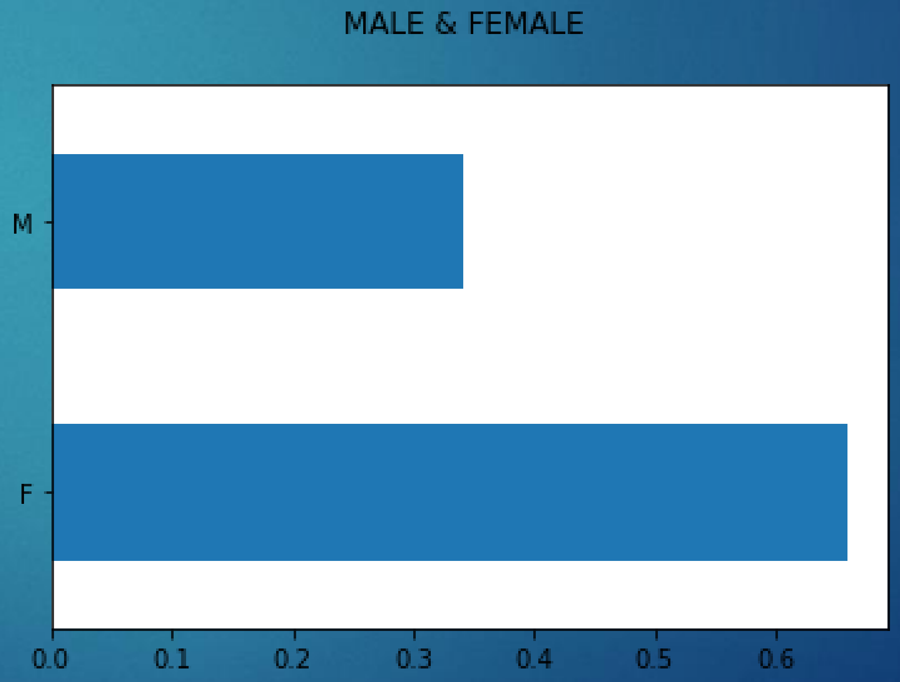
❑ Data Balancing

- I checked the data balance, In the data I check data describe 25%, 50%, and 75%.
- In order to make sure the columns were balanced, I removed them

❑ Analysis of Variables

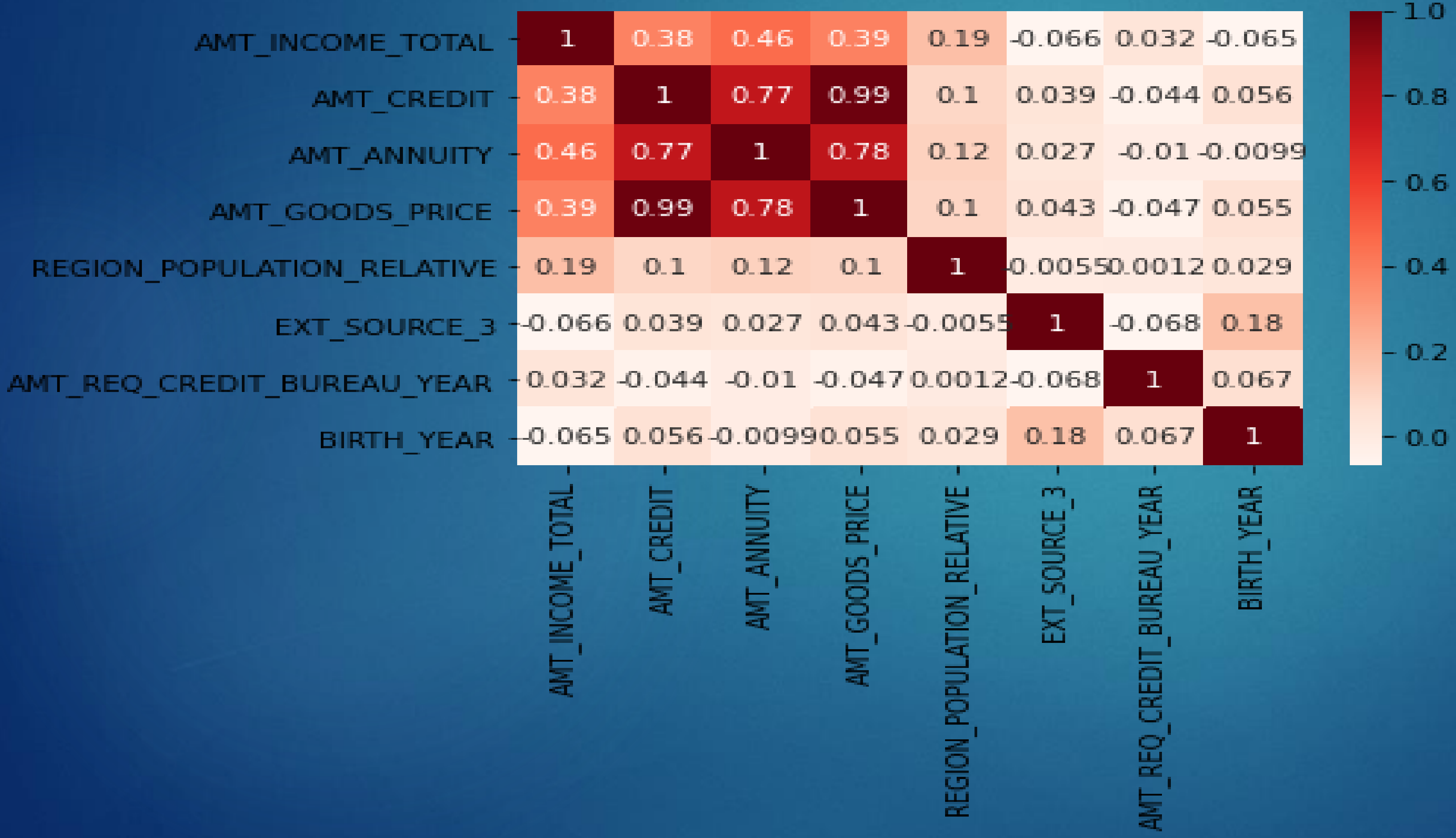
- A Barplot analysis has been performed on all columns


Target


MALE & FEMALE

❑ inside

- I checked with correlation plot how many columns are more related to each other



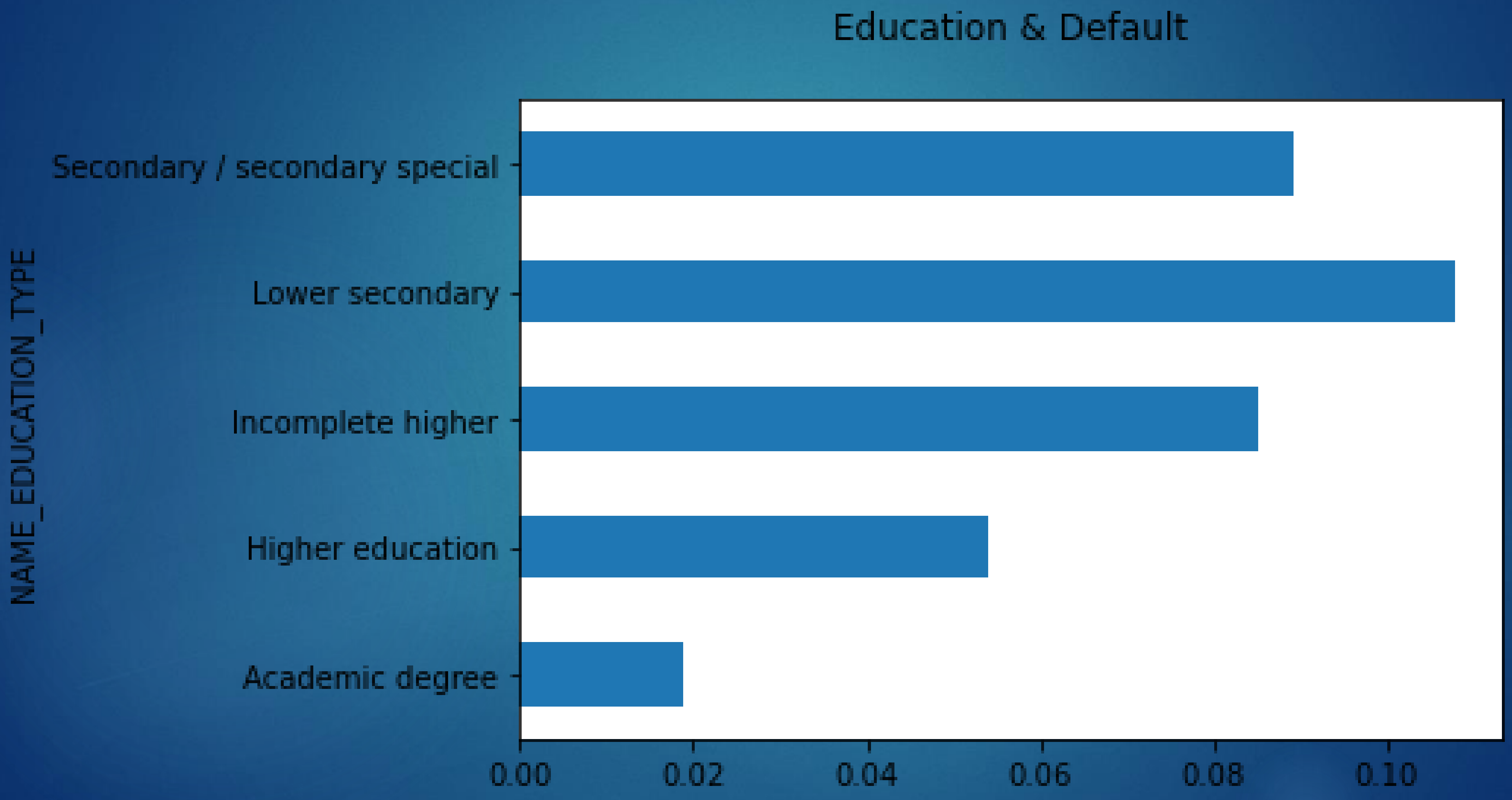
	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	EXT_SOURCE_3	AMT_REQ_CREDIT_BUREAU_YEAR	BIRTH_YEAR
AMT_INCOME_TOTAL	1	0.38	0.46	0.39	0.19	-0.066	0.032	-0.065
AMT_CREDIT	0.38	1	0.77	0.99	0.1	0.039	-0.044	0.056
AMT_ANNUITY	0.46	0.77	1	0.78	0.12	0.027	-0.01	-0.0099
AMT_GOODS_PRICE	0.39	0.99	0.78	1	0.1	0.043	-0.047	0.055
REGION_POPULATION_RELATIVE	0.19	0.1	0.12	0.1	1	-0.0055	0.0012	0.029
EXT_SOURCE_3	-0.066	0.039	0.027	0.043	-0.0055	1	-0.068	0.18
AMT_REQ_CREDIT_BUREAU_YEAR	0.032	-0.044	-0.01	-0.047	0.0012	-0.068	1	0.067
BIRTH_YEAR	-0.065	0.056	-0.0099	0.055	0.029	0.18	0.067	1

❑ inside

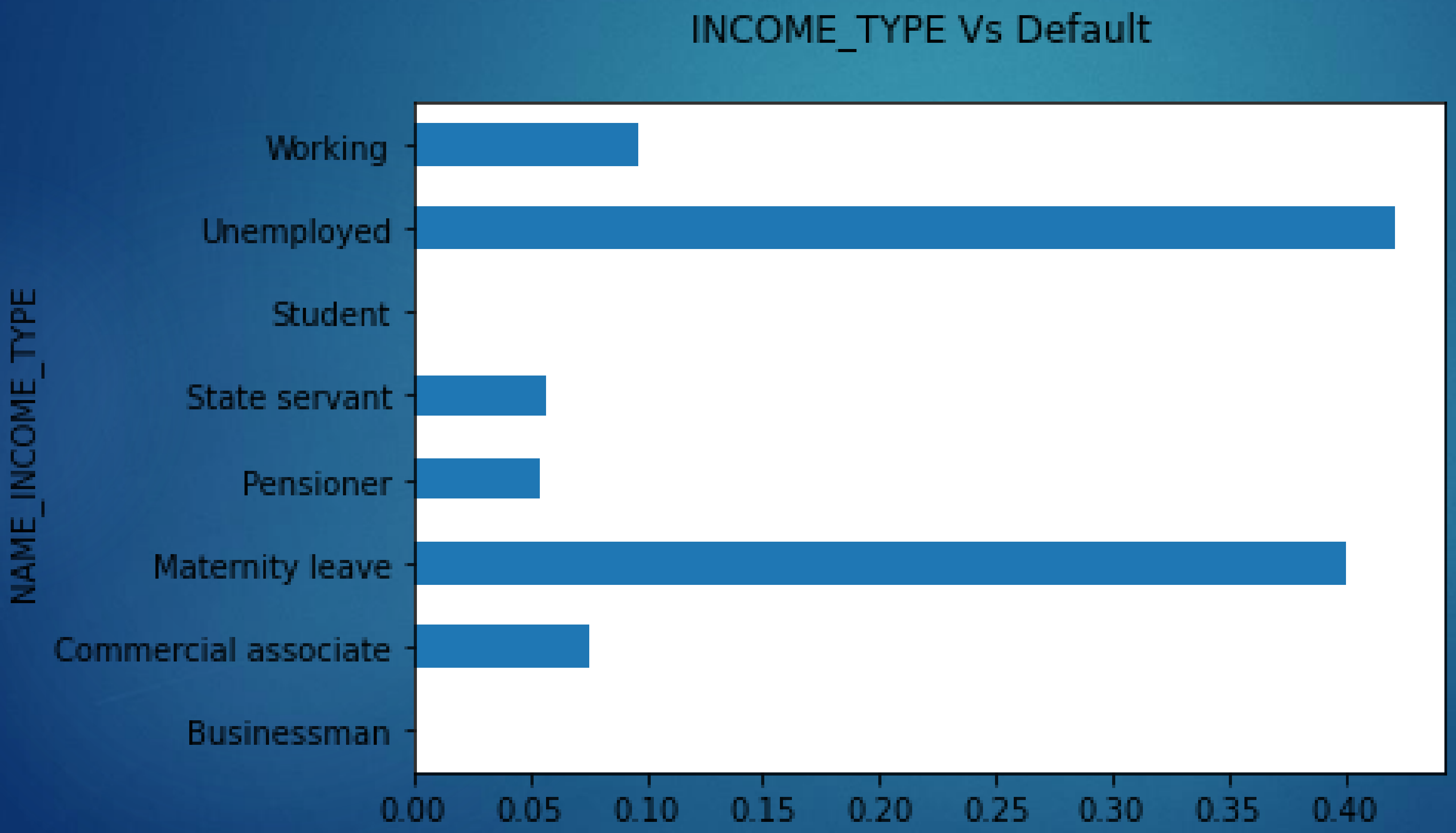
- As a group, I formed a column for education, and another for those who were unable to pay their loans, and plotted mean values on a bar graph

- And in that plot I have come to know that the customers with lower secondary education are not able to pay the loan much.

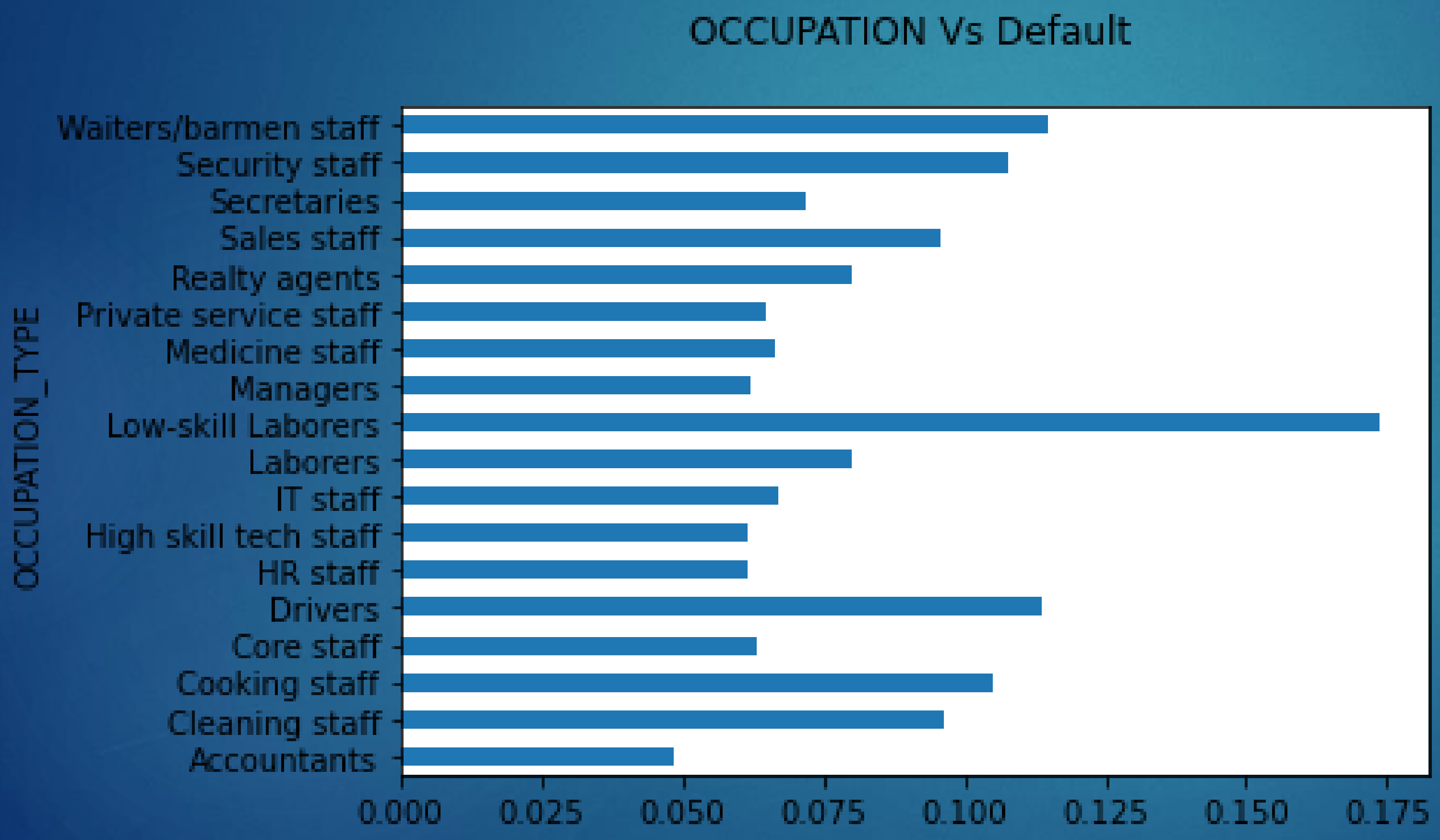
inside


Education & Default
NAME_EDUCATION_TYPE
Secondary / secondary special
Lower secondary
Incomplete higher
Higher education
Academic degree
0.00
0.02
0.04
0.06
0.08
0.10

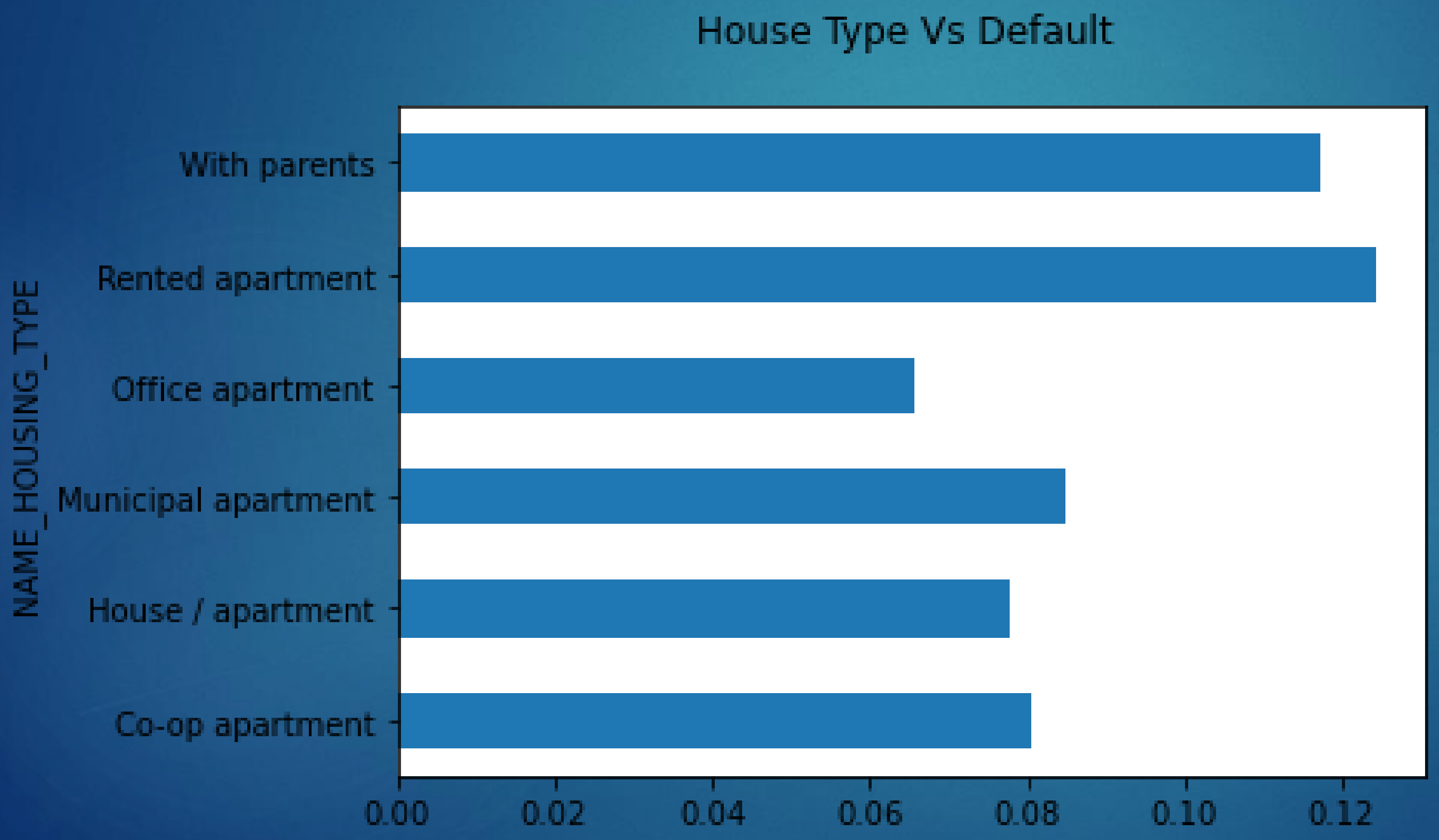
- Similarly, I plotted the income type and default customer, and I got the higher rate Maternity leave and Unemployed, unable to pay.



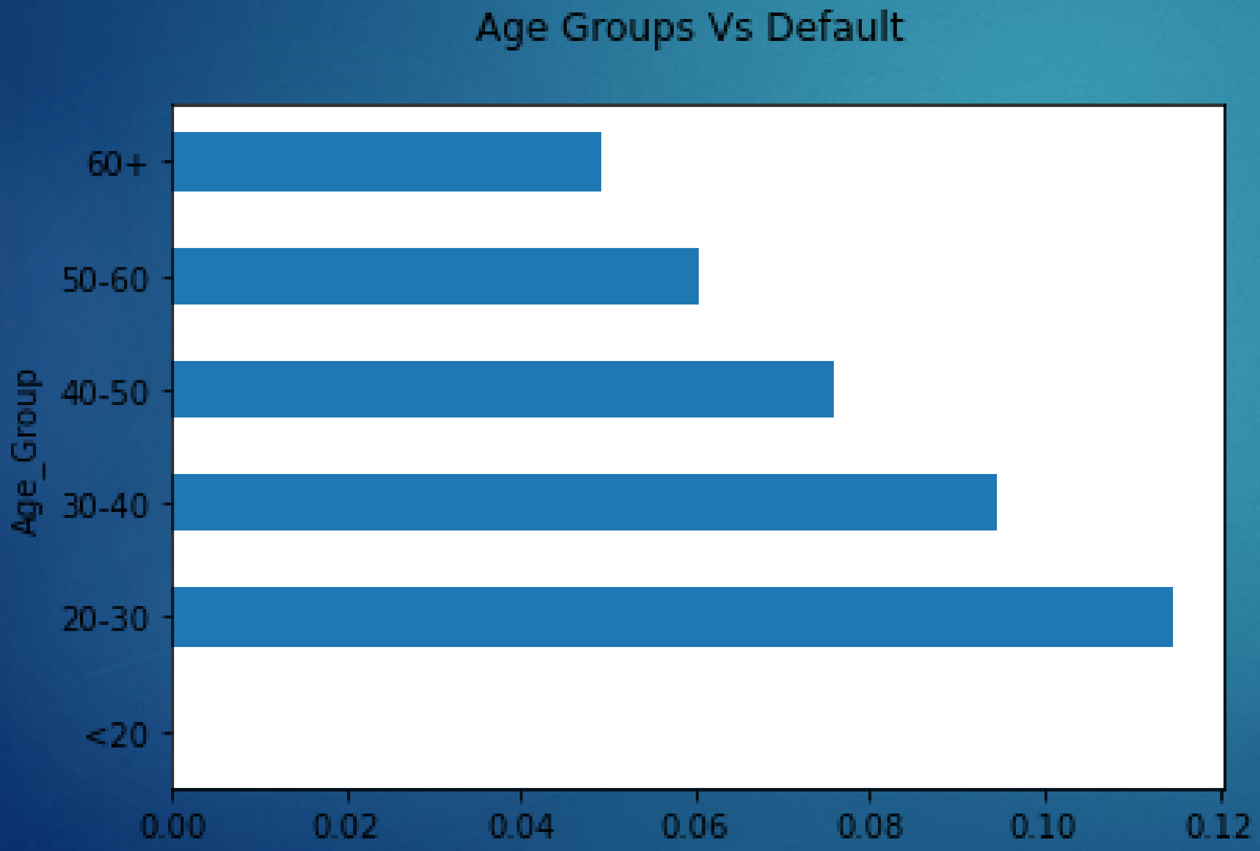
- Similarly, I plotted the OCCUPATION TYPE and default customer, and I got the higher rate Low-skill Laborers, unable to pay.



- Similarly, I plotted the HOUSING TYPE and default customer, and I got the higher rate in Rented apartment and With parents, unable to pay.



- First I split the age column in the difference bucket, and Similarly, I plotted the Age group and default customer, and I got the higher rate in between 20 to 30, unable to pay.


Age Groups Vs Default
Age_Group
<20