Credit EDA Assignment

Problem Statement:

In the given loan application data set, we are required to find factors/variables which influences default.

Steps followed for Analysis

1. Import all necessary libraries.

2. Load the application data set.

3. Get information about columns in the data set.

4. Create derived variables for income groups, age, age groups, employment_years, years_registration and years_id_publish for easy analysis.

4. Handling missing data

- Identify incorrect missing data notation in organization type and employment_years and replace it with NaN.
- Remove columns that has more than 40% missing data.

5. Check for outliers

- Outliers are data which are different and which do not fall into normal distribution of data.
- Outliers were identified in Amount_income_total, Amount_annuity, years_registration.

6. Data Imbalance

- The data set has huge number of records with target value 0(around 92%) compared to target value 1(around 8%). Hence the data set has data imbalance.

7. Perform Univariate Analysis

8. Perform Segmented Univariate Analysis

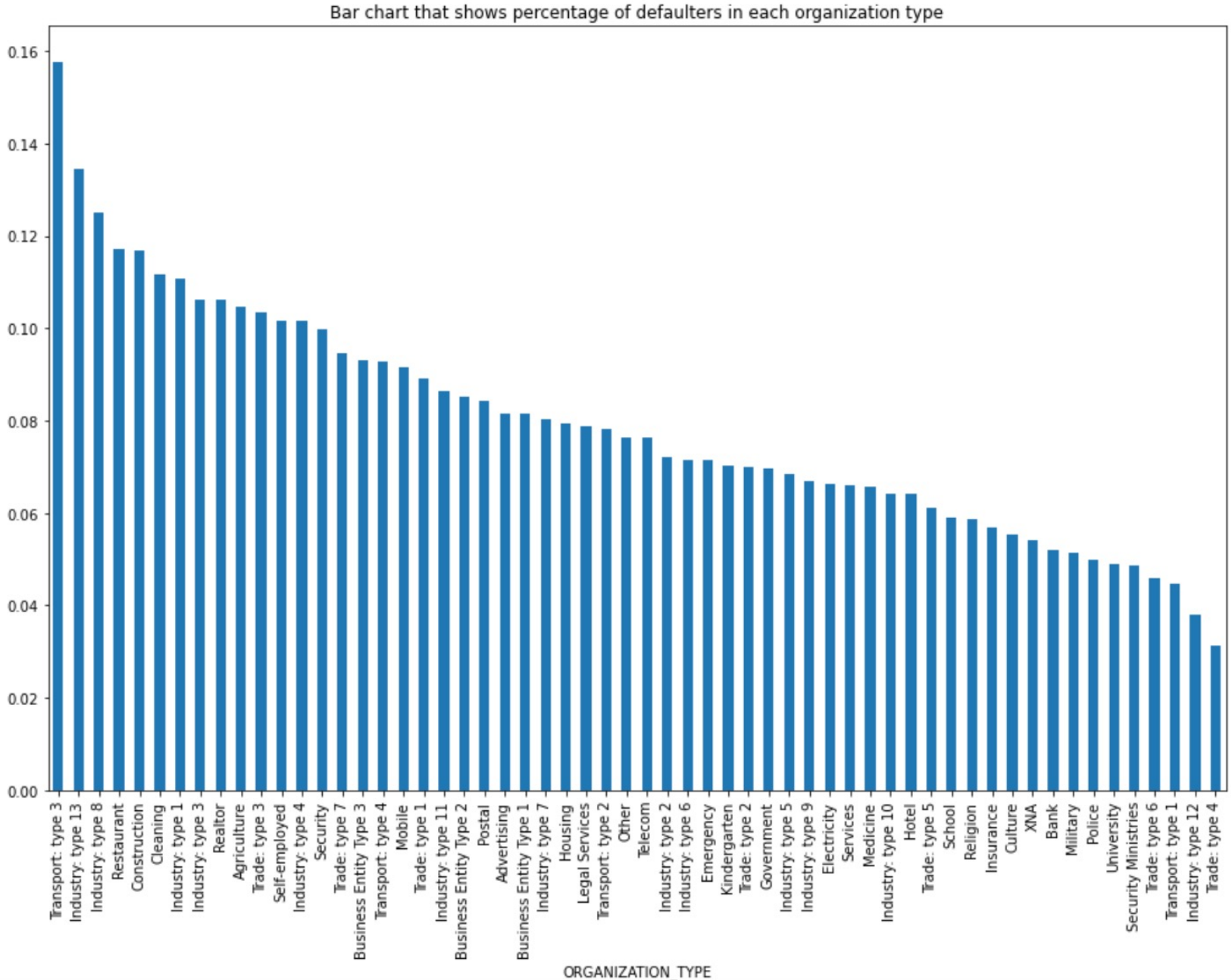
9. Perform Bivariate Analysis

Univariate Analysis

1. Organization Type

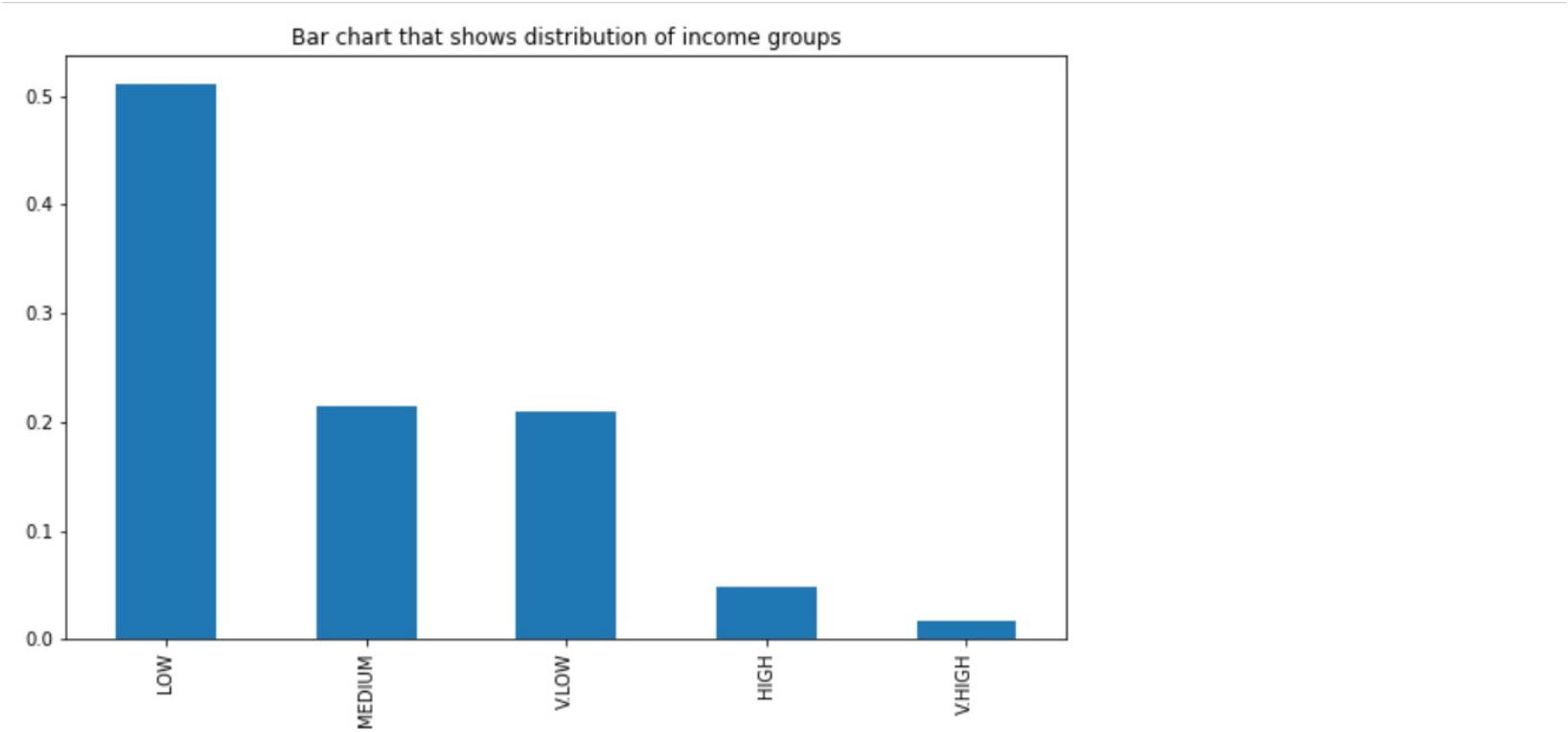
Though Univariate analysis done on organization type shows “Business Entity 3” as the major category

The percentage of defaulters is more in Transport Type 3 Category.


Bar chart that shows percentage of defaulters in each organization type
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Transport: type 3
Industry: type 13
Industry: type 8
Restaurant
Construction
Cleaning
Industry: type 1
Industry: type 3
Realtor
Agriculture
Trade: type 3
Self-employed
Industry: type 4
Security
Trade: type 7
Business Entity Type 3
Transport: type 4
Mobile
Trade: type 1
Industry: type 11
Business Entity Type 2
Postal
Advertising
Business Entity Type 1
Industry: type 7
Housing
Legal Services
Transport: type 2
Other
Telecom
Industry: type 2
Industry: type 6
Emergency
Kindergarten
Trade: type 2
Government
Industry: type 5
Industry: type 9
Electricity
Services
Medicine
Industry: type 10
Hotel
Trade: type 5
School
Religion
Insurance
Culture
XNA
Bank
Military
Police
University
Security Ministries
Trade: type 6
Transport: type 1
Industry: type 12
Trade: type 4
ORGANIZATION_TYPE

2. Amount Income Group

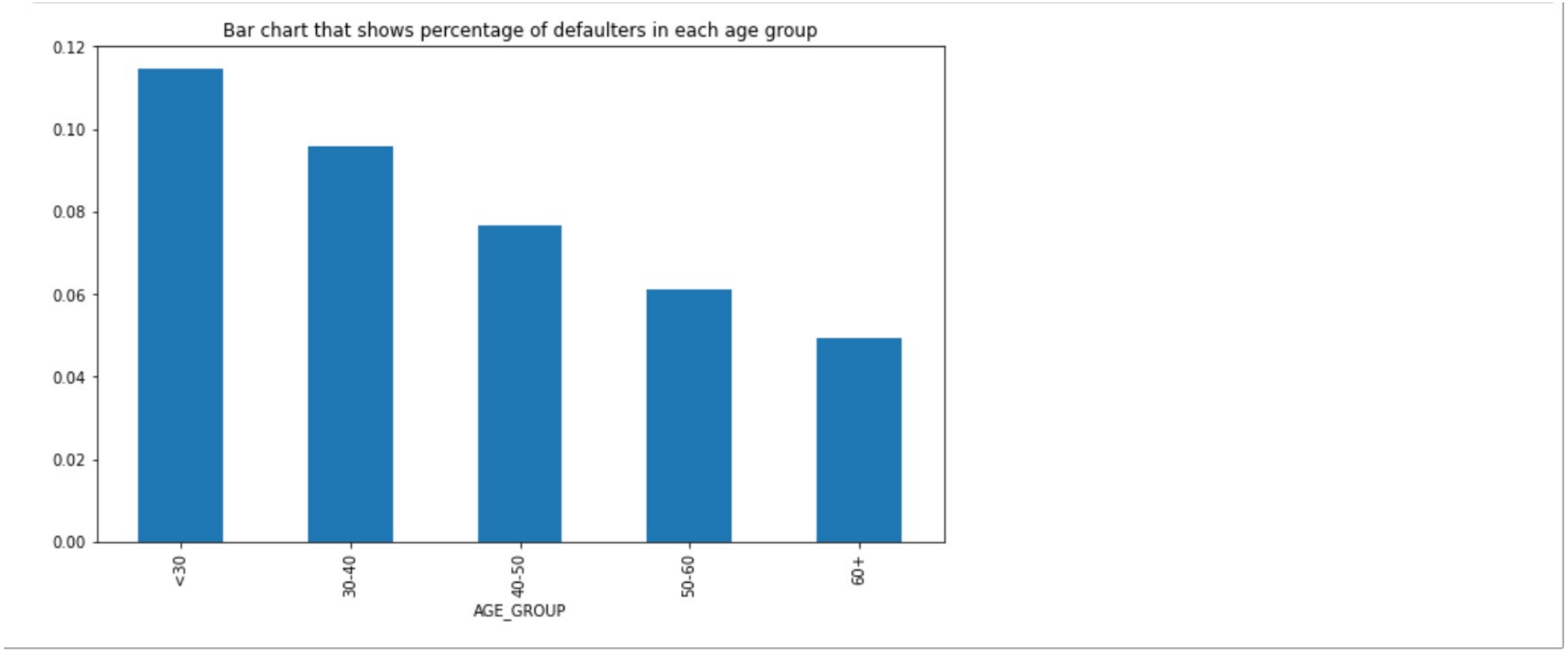
Univariate analysis done on Amount Income Group shows income category “Low” forms the majority in the data set.



From above graph, we observe majority of people in the data set fall under Low income category

3. Age Group

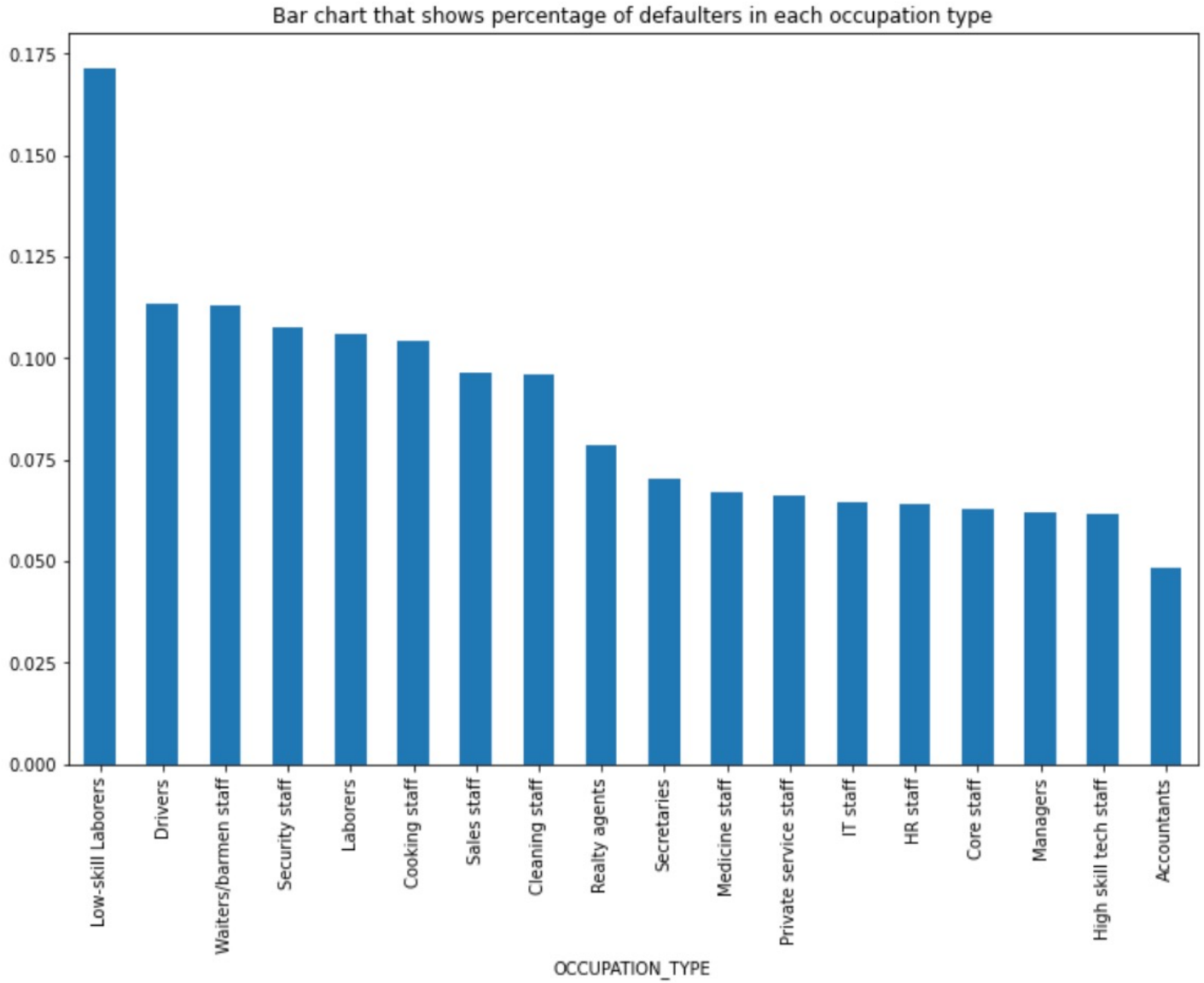
Though Univariate analysis done on Age Group shows 30-40 age group category as majority. Percentage of defaulters is more in <30 age group category.



Though our data set contains majority of people in 30-40 age group category. The percentage of defaulters is most in <30 age group.

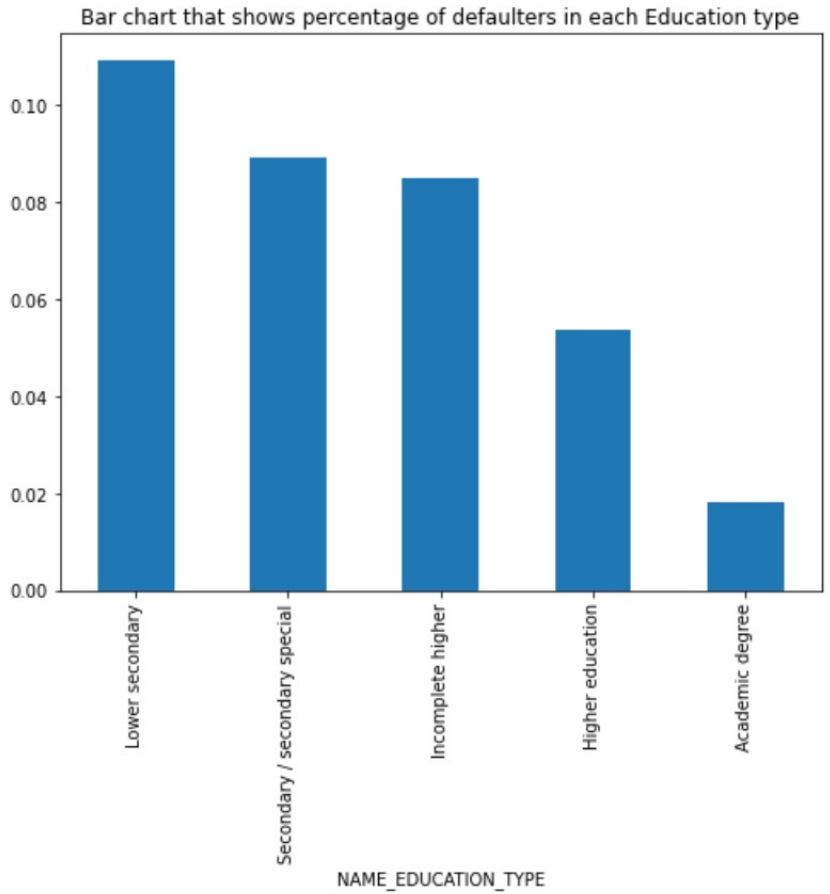
4. Occupation Type

Though Univariate analysis done on Occupation Type shows “Laborers” as majority. Percentage of defaulters is more in Low-Skilled Laborers.



5. Education Type

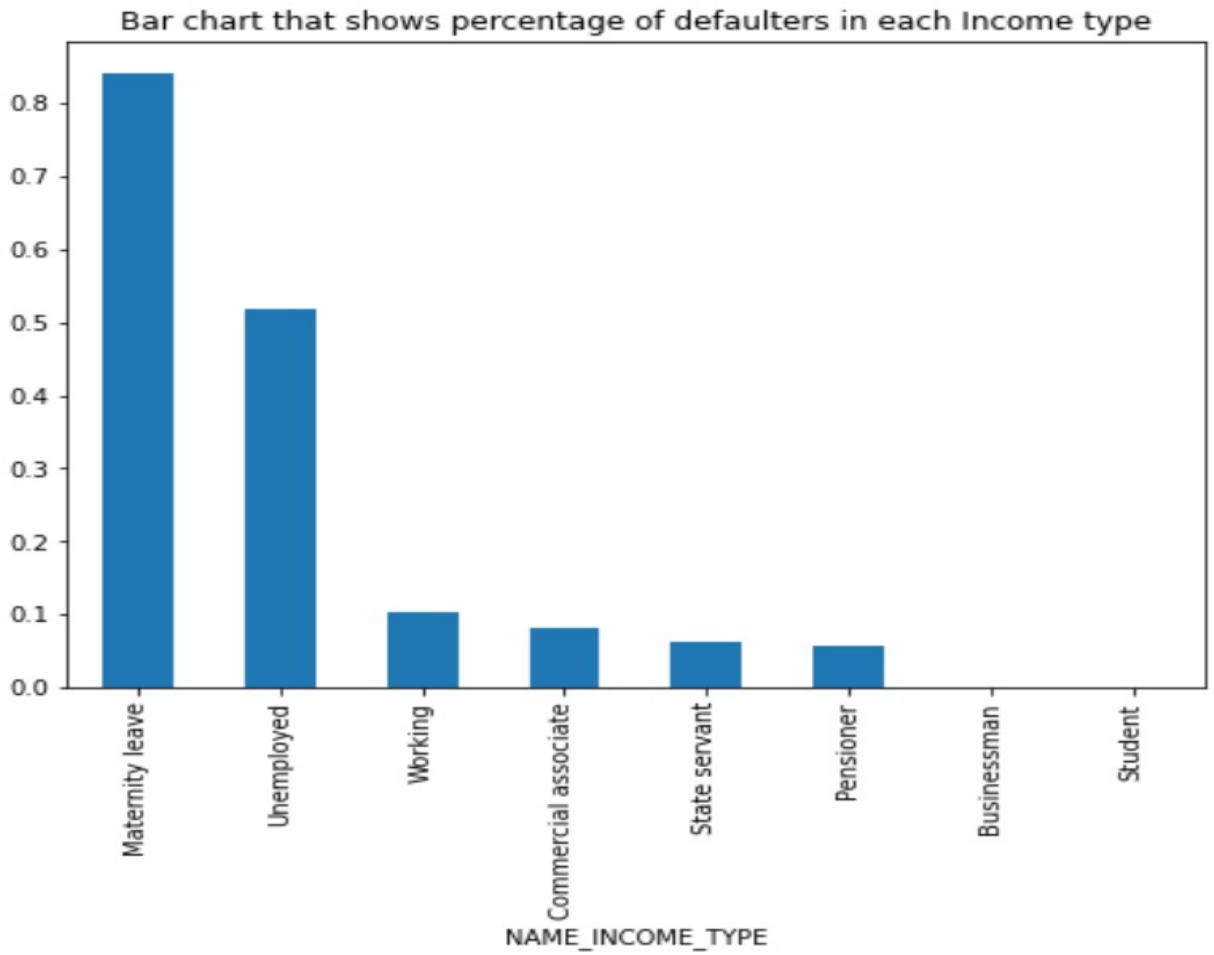
Though Univariate analysis done on Education Type shows “Secondary” as majority. Percentage of defaulters is more in Lower Secondary category.



Though majority of people in the data set belong to secondary education category. The propotion of defaulters is more in Lower secondary category.

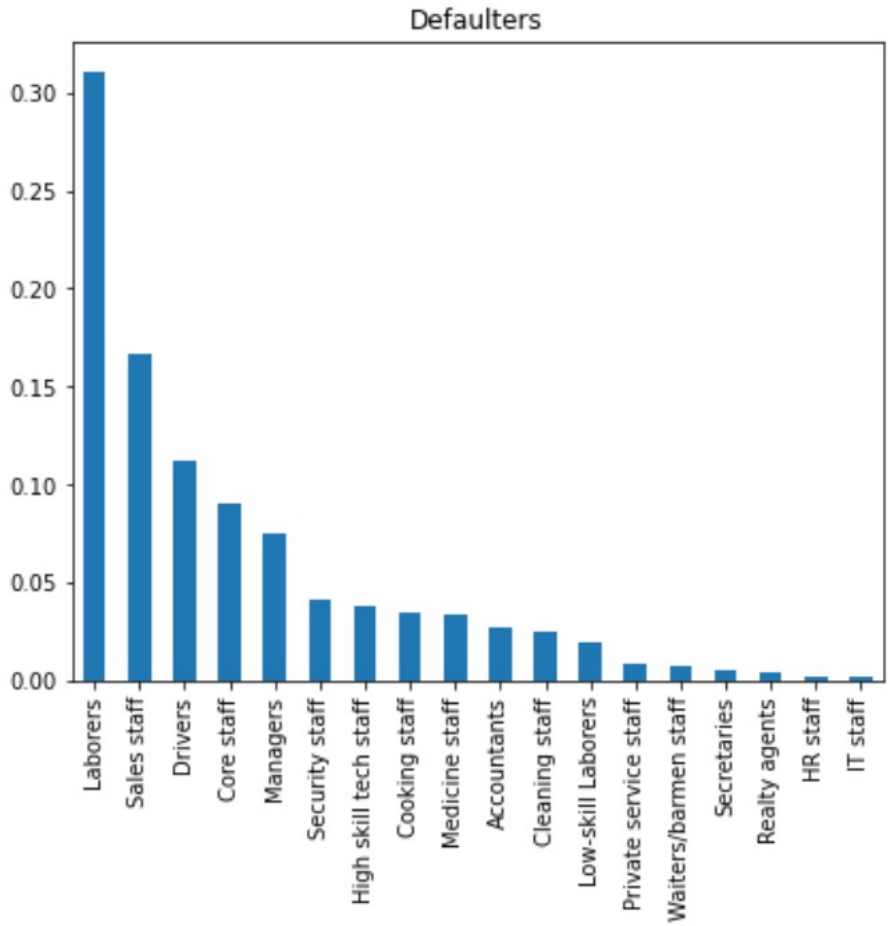
6. Income Type

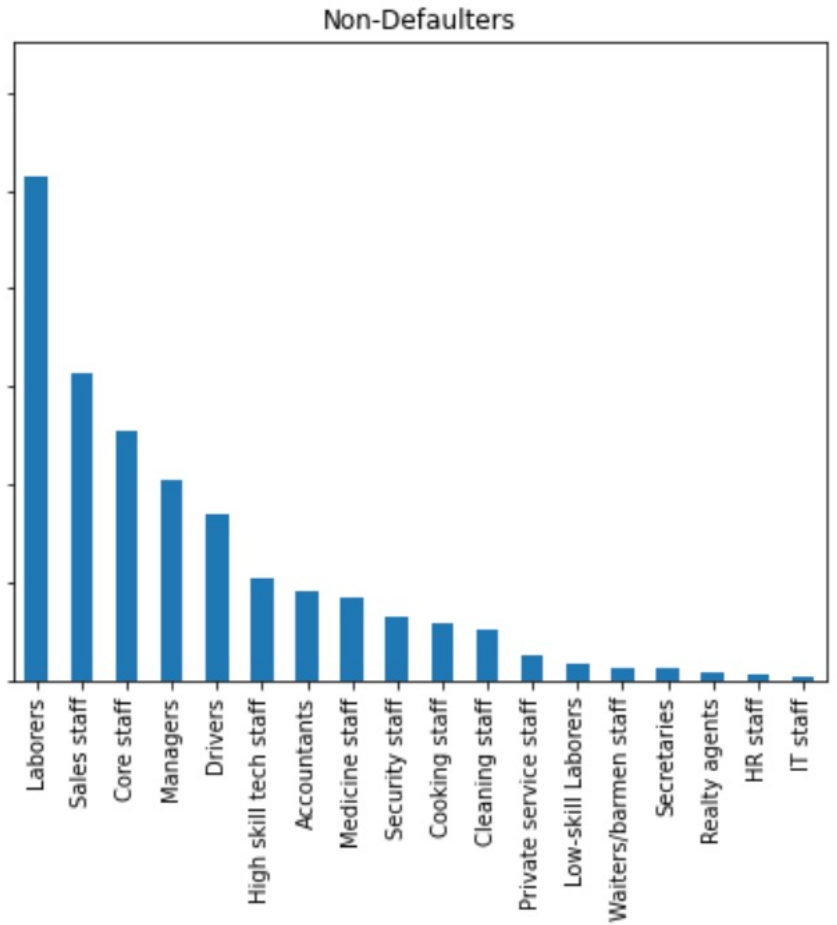
Though Univariate analysis done on Income Type shows “Working” as majority. Percentage of defaulters is more in Maternity leave category.



Segmented Univariate Analysis

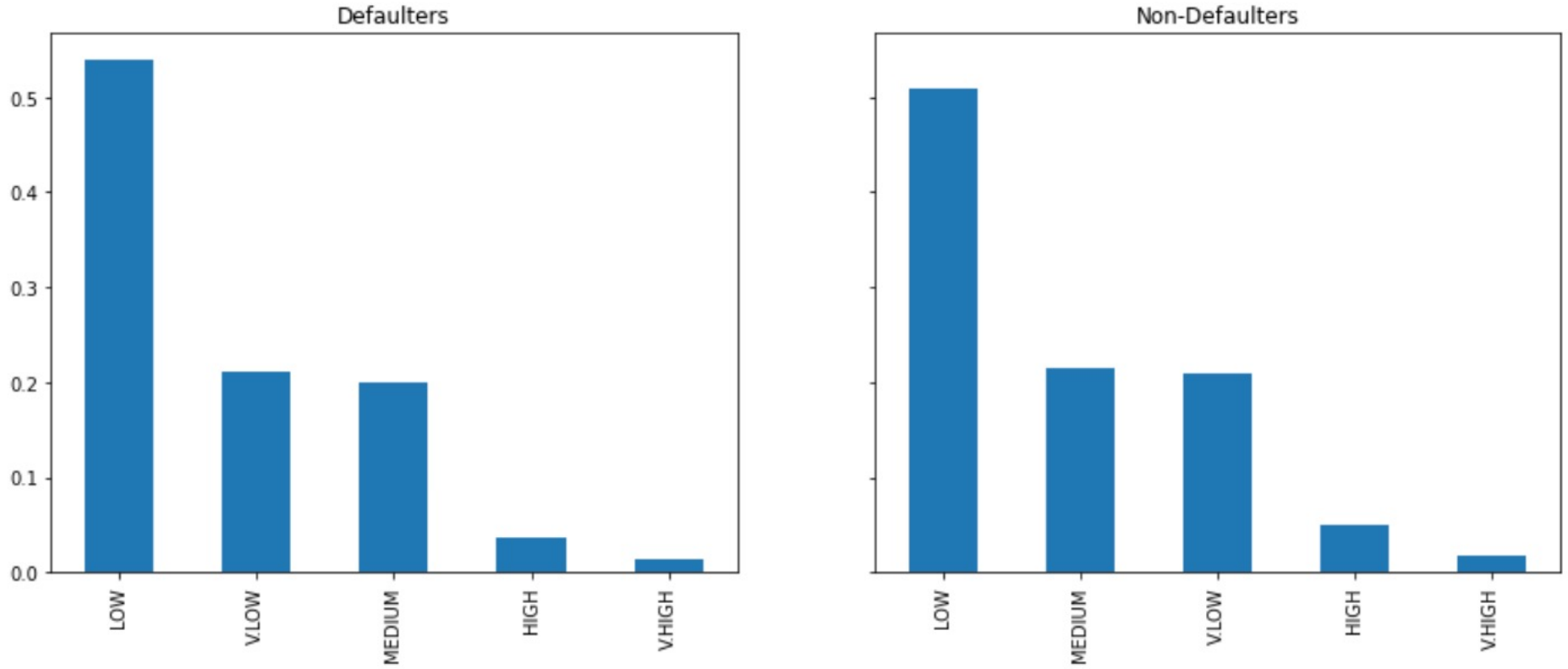
- Distribution of Occupation Type for Defaulters and Non-Defaulters





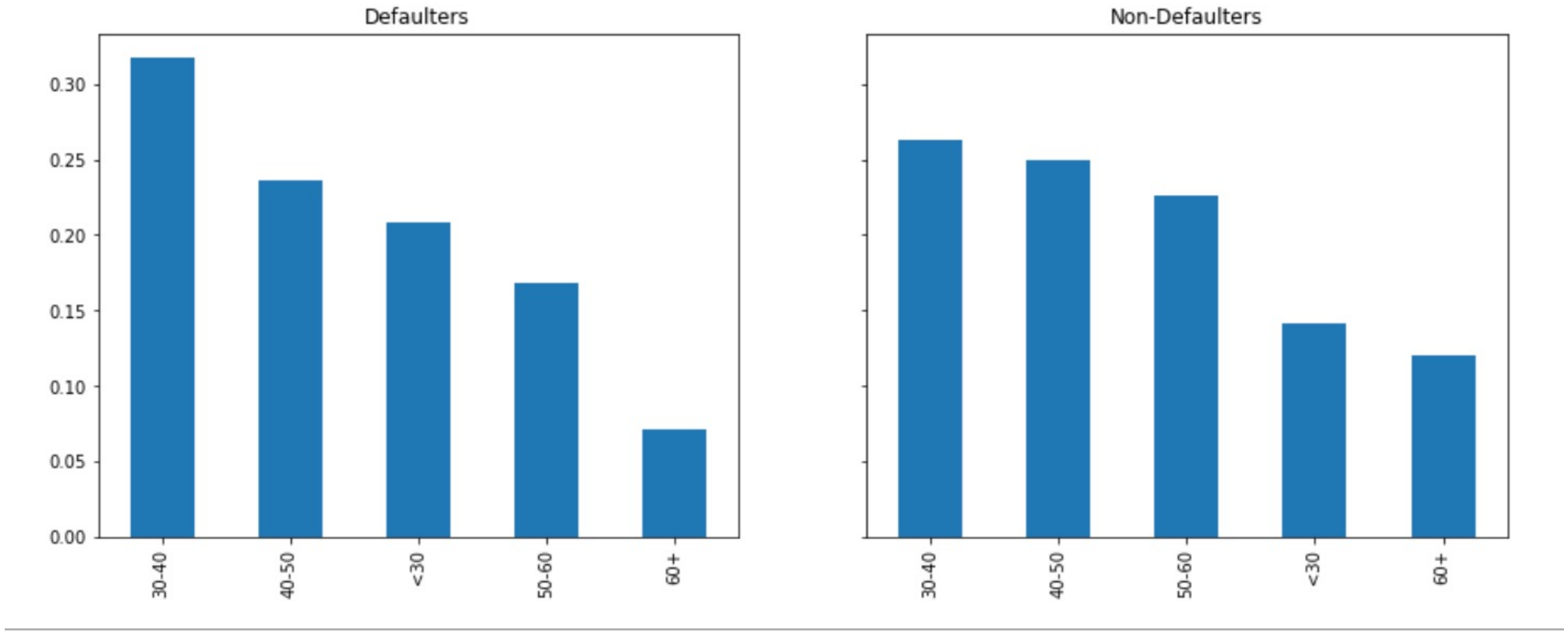
In Labourers occupation type, propotion of defaulters than non-defaulters is high

- Distribution of Income group for Defaulters and Non-Defaulters



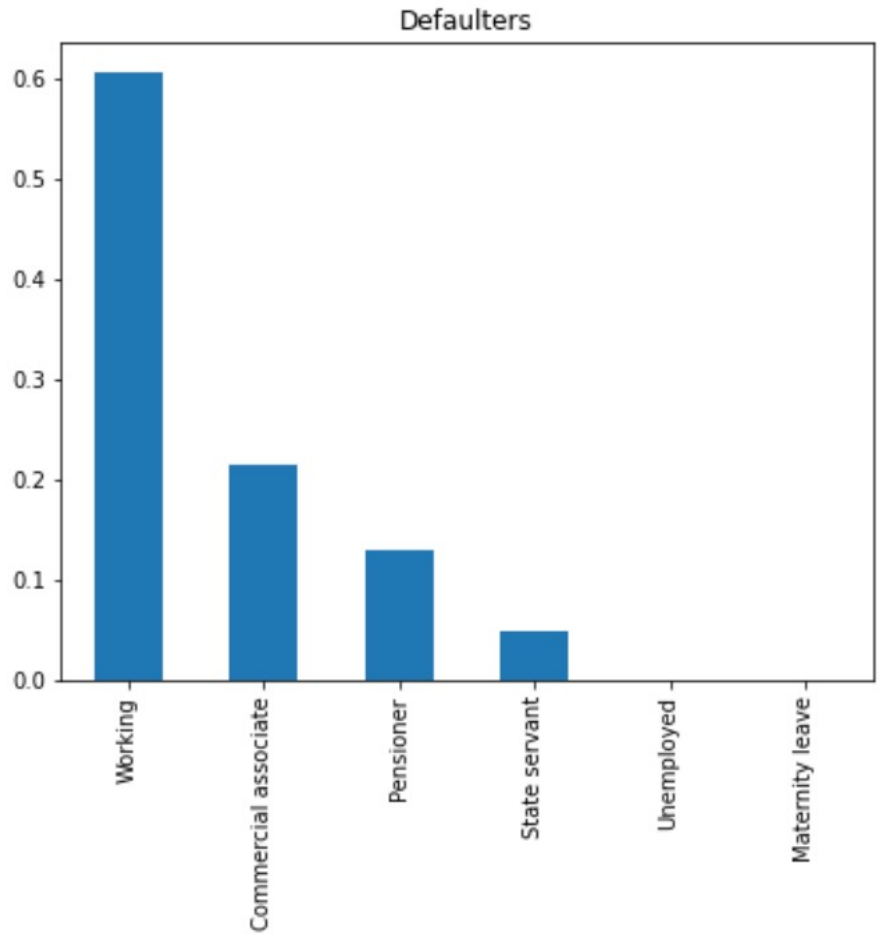
In low income group, propotion of defaulters than non-defaulters is high

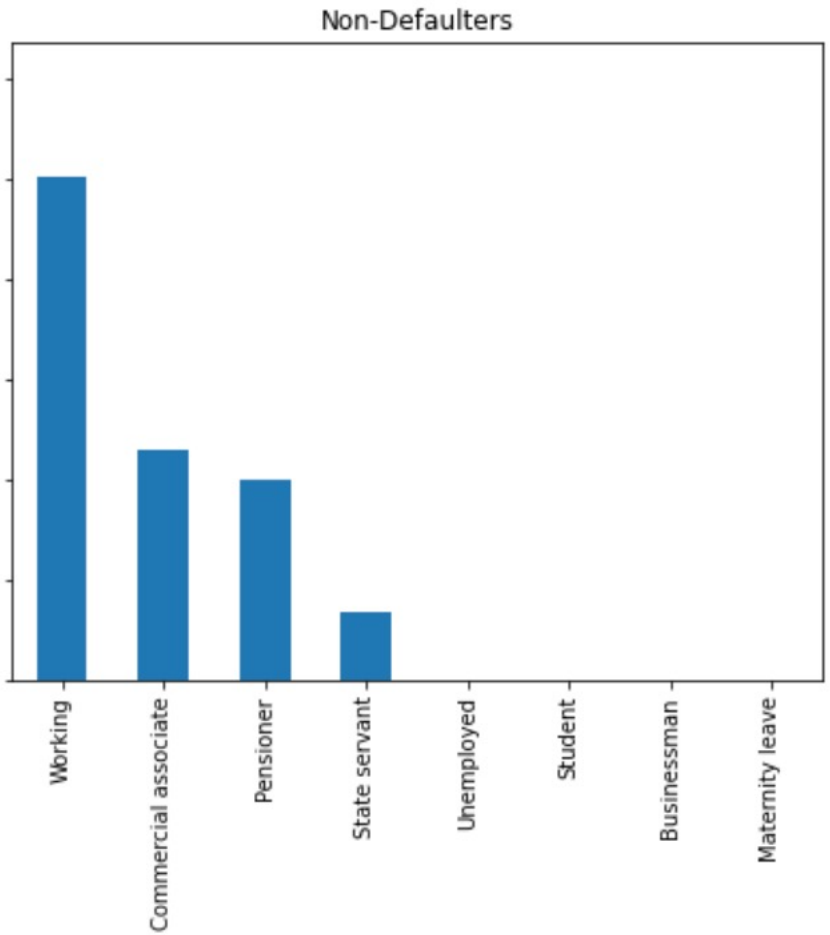
- Distribution of Age group for Defaulters and Non-Defaulters



In 30-40 age group category, propotion of defaulters is higher than non-defaulters

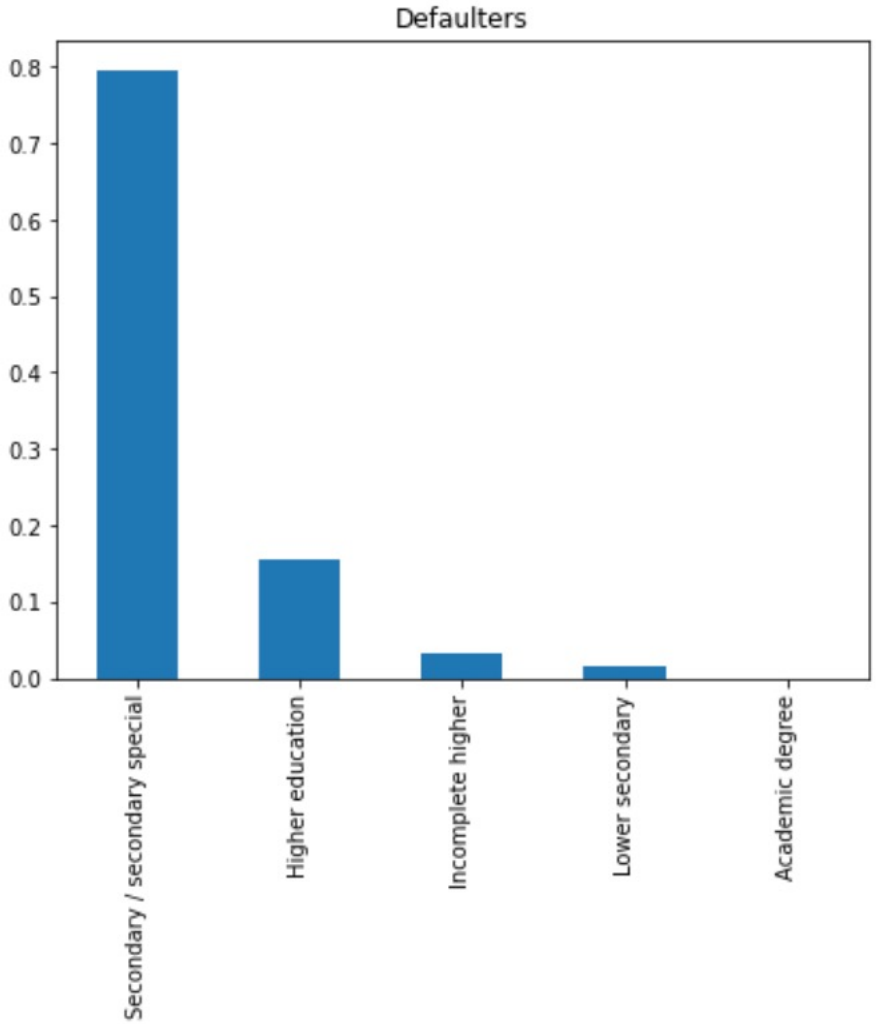
- Distribution of Income Type for Defaulters and Non-Defaulters

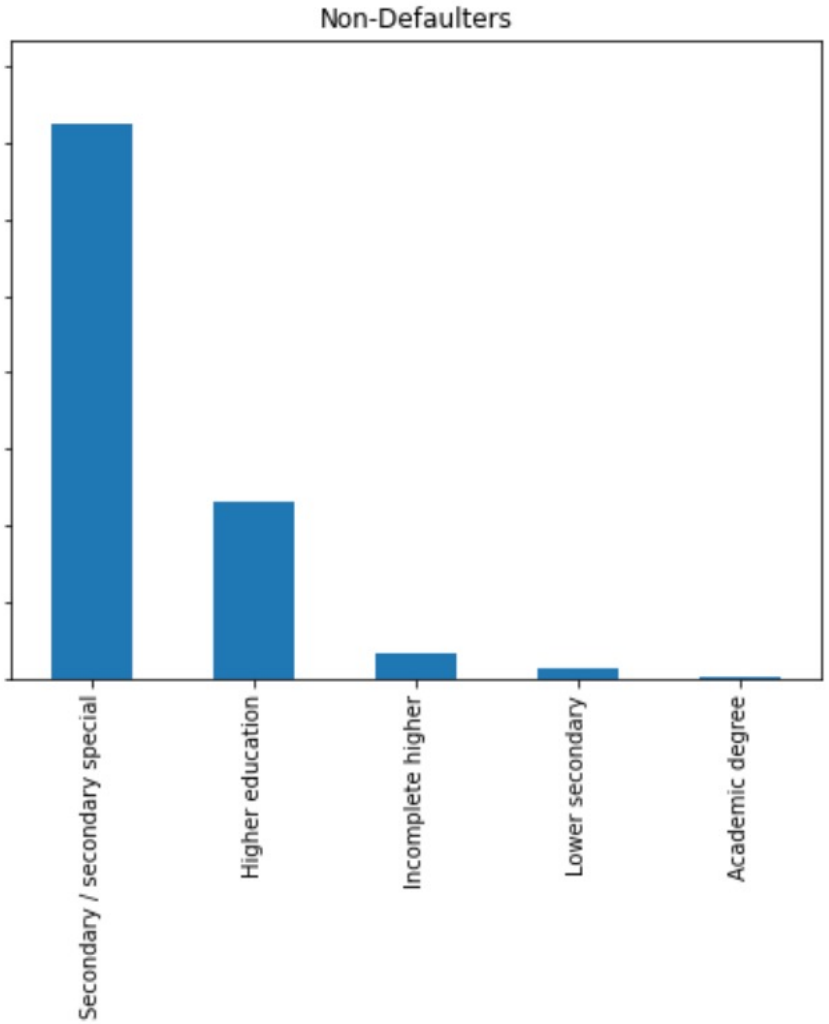




In Working Income Type category, propotion of defaulters is higher than non-defaulters

- Distribution of Education Type for Defaulters and Non-Defaulters





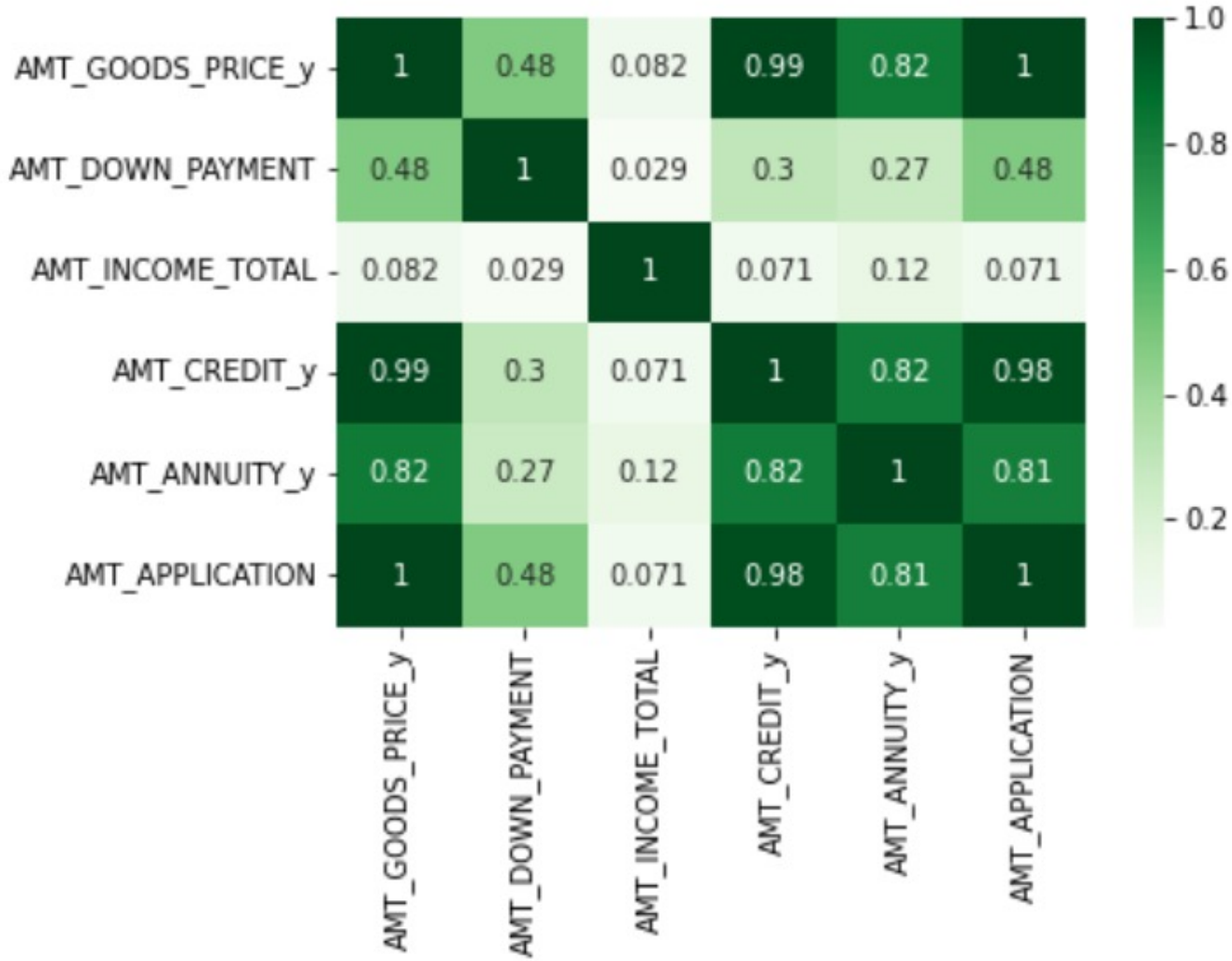
In Secondary Education Type category, propotion of defaulters is higher than non-defaulters

Bivariate Analysis

- In current application data set, strong correlation is absorbed between the following variables.
- Strong correlation means when one variable increases, the other variable increases too.
- 1. AMT_CREDIT and AMT_ANNUITY,
- 2. AGE_GROUP and EMPLOYED_YEARS,
- 3. AMT_GOODS_PRICE and AMT_CREDIT,
- 4. AMT_GOODS_PRICE and AMT_ANNUITY,
- 5. AGE and YEARS_REGISTRATION,
- 6. AGE and YEARS_ID_PUBLISH
- It can be seen through the following heat map.


AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
EMPLOYED_YEARS
AGE
YEARS_REGISTRATION
YEARS_ID_PUBLISH
1 0.16 0.19 0.16 0.013 -0.027 -0.028 -0.0085
0.16 1 0.77 0.99 0.091 0.055 -0.0096 0.0066
0.19 0.77 1 0.78 0.054 -0.0094 -0.039 -0.011
0.16 0.99 0.78 1 0.093 0.053 -0.012 0.0093
0.013 0.091 0.054 0.093 1 0.35 0.17 0.09
-0.027 0.055 -0.0094 0.053 0.35 1 0.33 0.27
-0.028 -0.0096 -0.039 -0.012 0.17 0.33 1 0.1
-0.0085 0.0066 -0.011 0.0093 0.09 0.27 0.1 1
1.0
0.8
0.6
0.4
0.2
0.0

- In the previous application data, there is strong correlation between
- 1. AMT_GOODS_PRICE and AMT_CREDIT,
- 2. AMT_GOODS_PRICE and AMT_ANNUITY,
- 3. AMT_DOWN_PAYMENT and AMT_GOODS_PRICE,
- 4. AMT_APPLICATION and AMT_CREDIT
- It can be seen through the following heat map.


AMT_GOODS_PRICE_y
AMT_DOWN_PAYMENT
AMT_INCOME_TOTAL
AMT_CREDIT_y
AMT_ANNUITY_y
AMT_APPLICATION
1 0.48 0.082 0.99 0.82 1
0.48 1 0.029 0.3 0.27 0.48
0.082 0.029 1 0.071 0.12 0.071
0.99 0.3 0.071 1 0.82 0.98
0.82 0.27 0.12 0.82 1 0.81
1 0.48 0.071 0.98 0.81 1
AMT_GOODS_PRICE_y
AMT_DOWN_PAYMENT
AMT_INCOME_TOTAL
AMT_CREDIT_y
AMT_ANNUITY_y
AMT_APPLICATION
1.0
0.8
0.6
0.4
0.2

Conclusion

In the given data set, people belonging to the following groups have more chances of default.

1. Low income group
2. Age group <30
3. Low Skilled Laborers occupation type.
4. Transport Type 3 organization Type.
5. Lower Secondary Education type.
6. Maternity leave working type.