EDA ASSIGNMENT ON BANK DATA SET

SUBMITTED BY

SUBHAM DEY

BATCH – DS – C38

PROGRAM : upGrad & IIITB | Data Science Program – November 2021

Data : Banking and finance loan service data set

Objective :

- To identify a pattern which indicates if a client has difficulty in loan payments.
- Key parameters that can affect the loan repayment

Data files provided :

- '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.
- '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.
- 'columns_description.csv' is data dictionary which describes the meaning of the variables.

Understanding & cleaning of data:

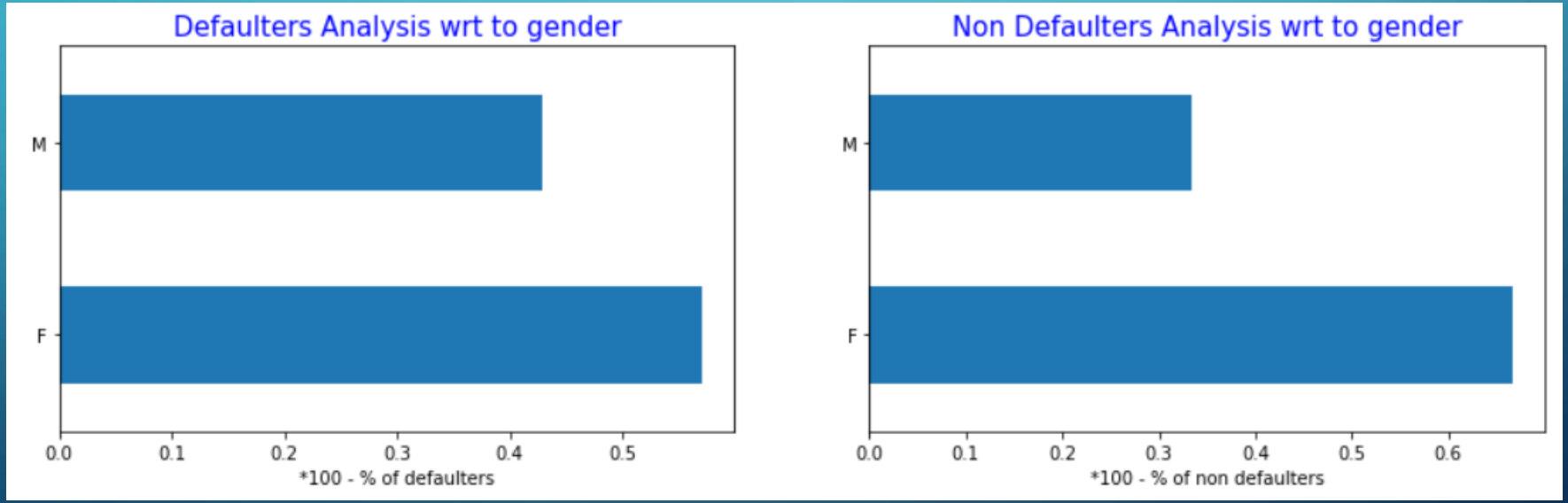
- Necessary checks are done on the data using functions like info(), describe(), .shape, .head() etc on both the data set.
- Checked for presence of duplicated data
- Data cleaning is then applied to the data set
- Check for NA values and necessary action is then taken on those
- Imputation is done on some of the column features
- Check for correct data types and standardizing data
- Check for outliers and dealt with them

Analysis of data:

- As the Target column indicates the defaulters and non defaulters we just need to separate the 2 data frames one with 1's and another with 0's

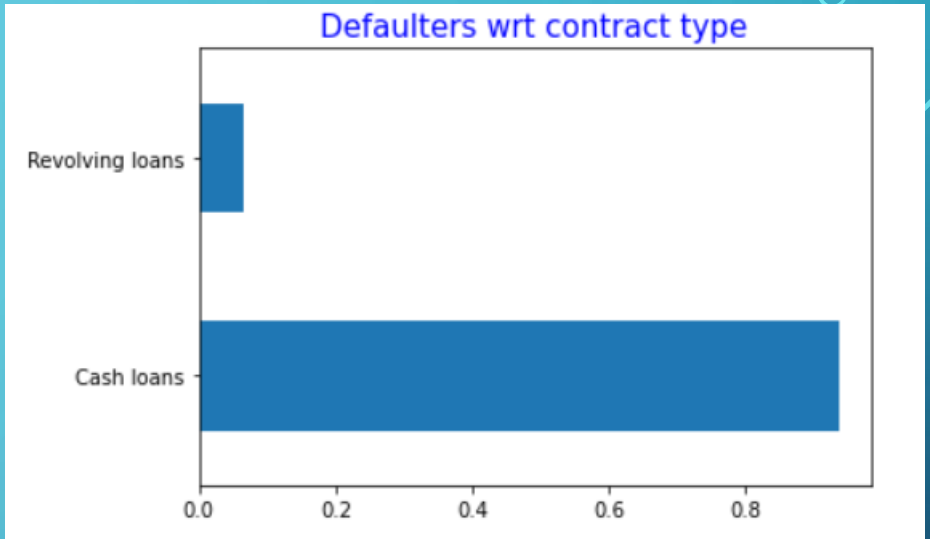
Univariate analysis (Categorical)

- The pattern for percentage of gender for both defaulter and non defaulter looks like same female candidates are more in both the cases

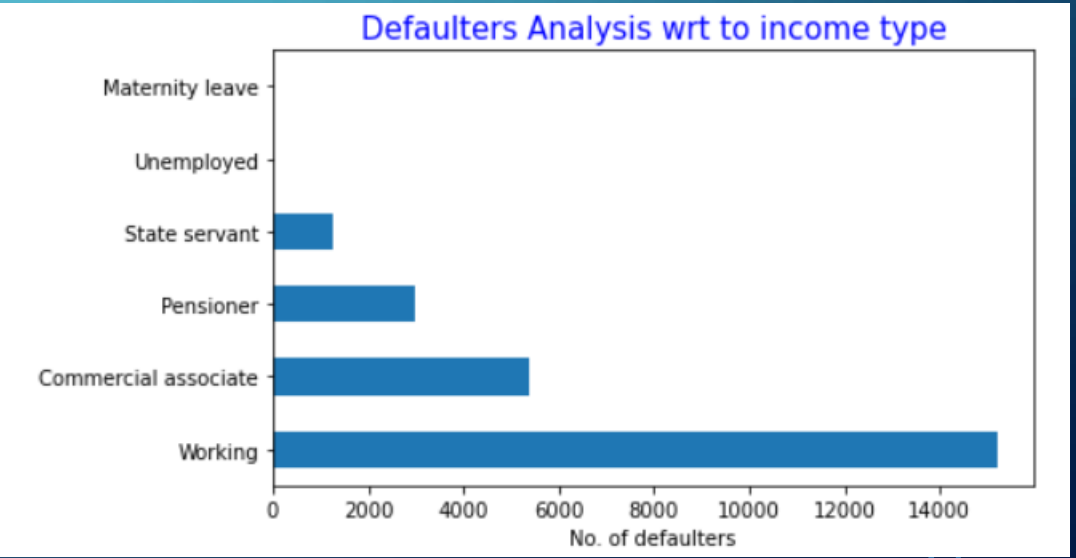


Univariate analysis (Categorical)

- From the figure we can see almost 90% of the clients who have taken cash loan are having repayment issues.

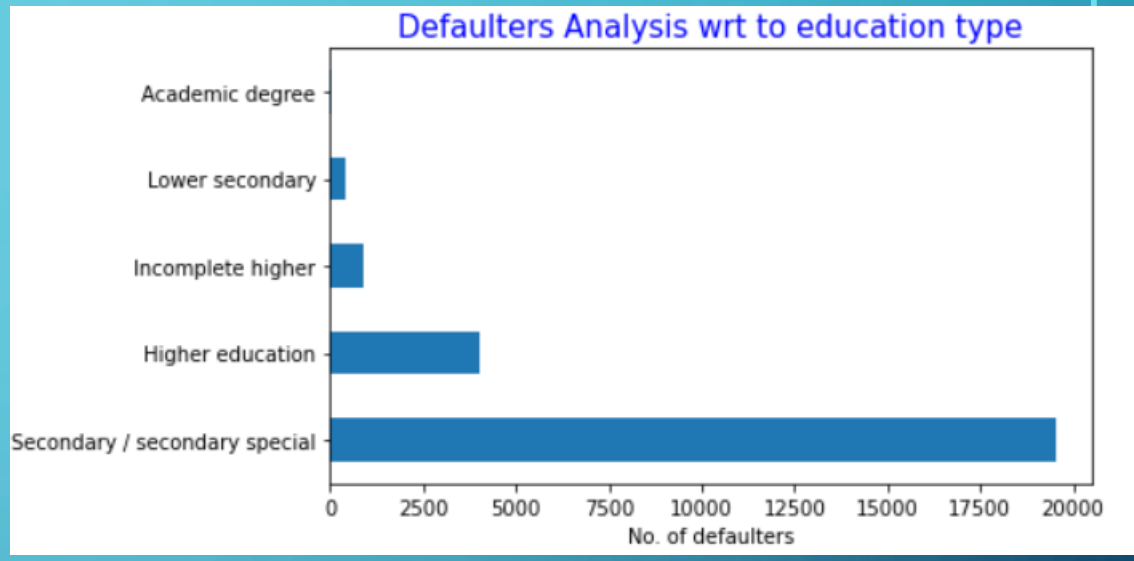


- There is an increasing pattern of defaulters for the income type, numbers are growing high from unemployed to working

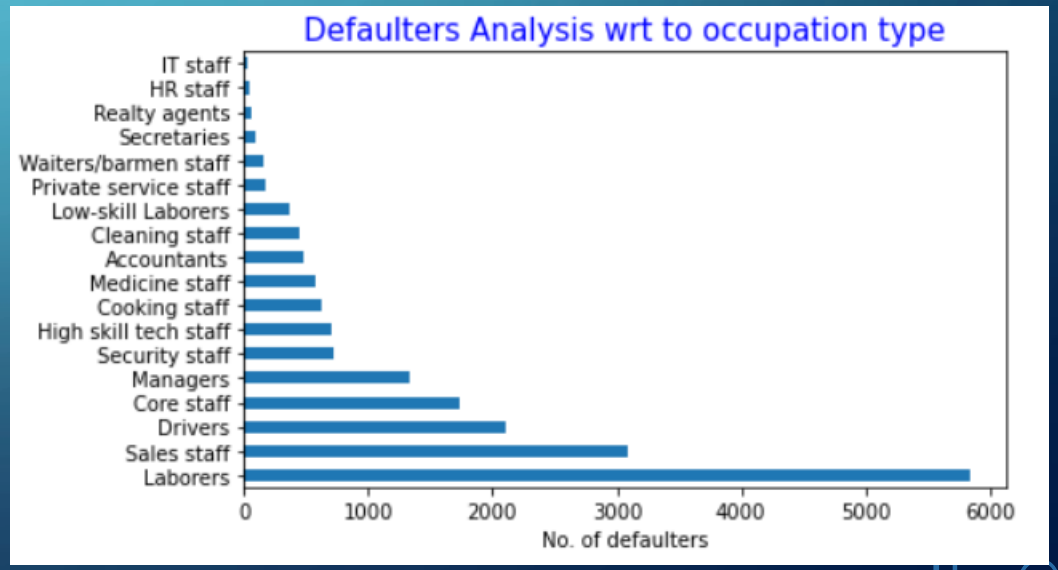


Univariate analysis (Categorical)

- An increasing pattern is followed in this case
- People with Secondary education are likely have more no. of defaulters compared to other categories

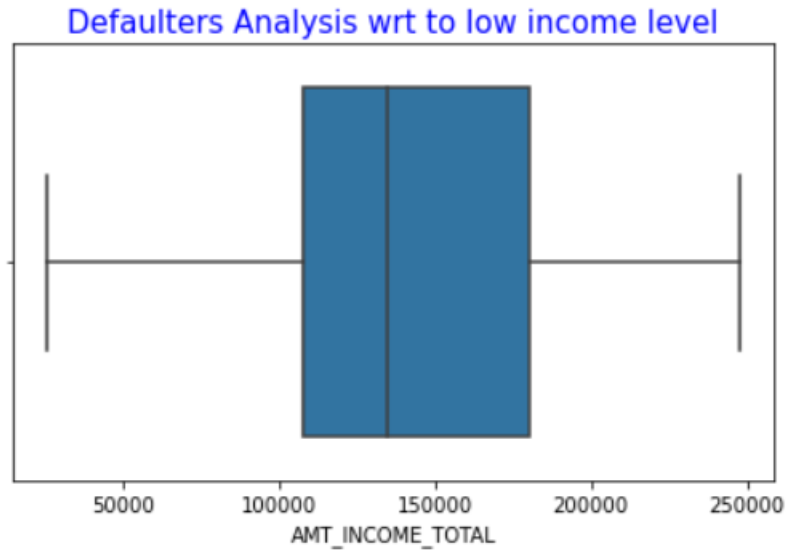


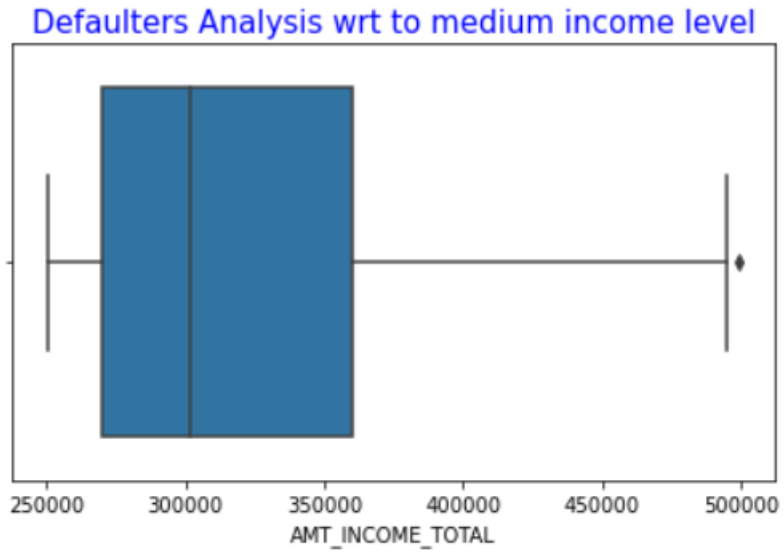
- This can be observed that IT occupations are less likely to become a defaulter compared to other categories
- Labours are having very high chances to do the repayment.
- High skill tech and security staffs are having almost similar affects.



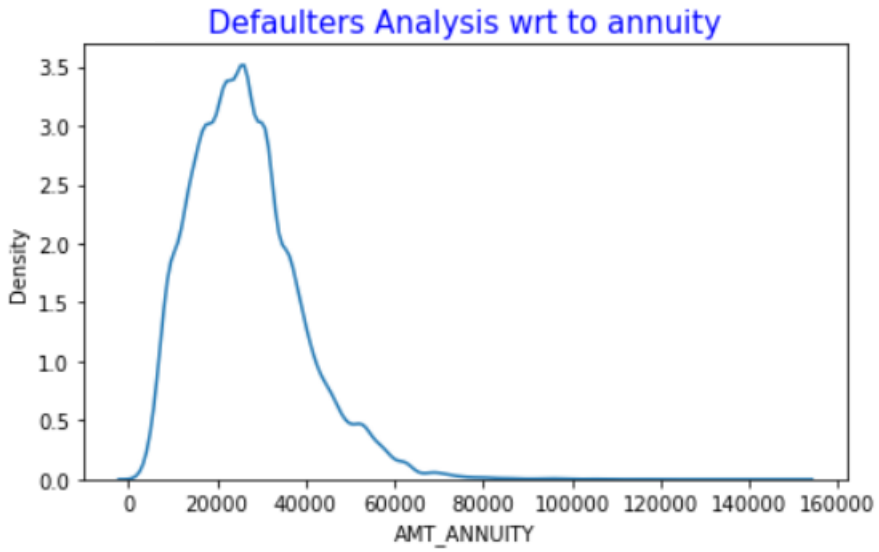
Univariate analysis (Numerical)

- A Whisker plot is created on the low income level (assuming Low = Income <= 2.5L, Medium = 5L<Income<=15L, High = Income>15L)
- Clients with low salary in-between 108000 to 180000 are having higher chances of loan repayment issue
- Medium salary in-between 270000 to 360000 are having higher chances of loan repayment issue

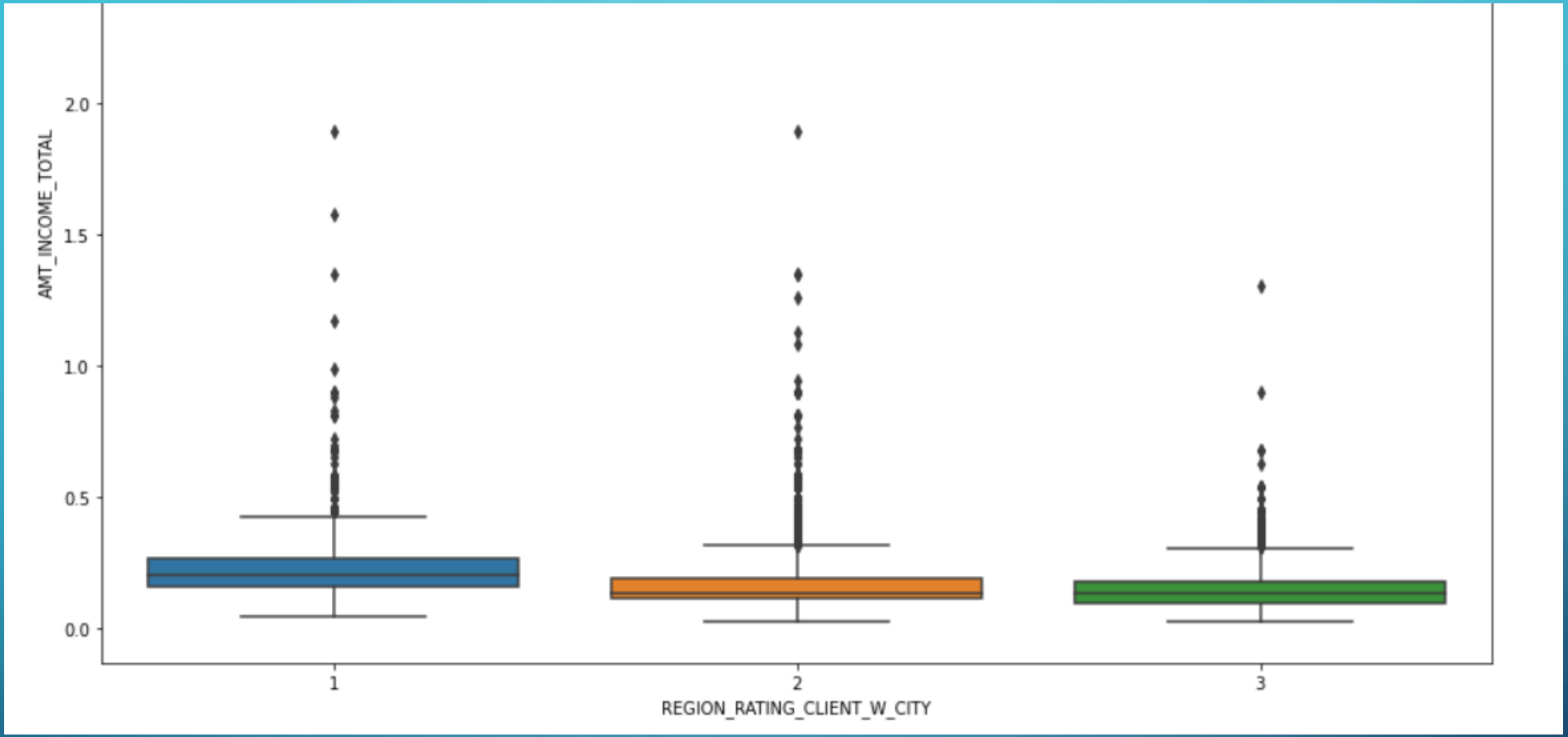




- It can be observed that higher annuity amount has lower payment difficulties but higher annuity amount cannot be suggested for reducing defaulters in loan.



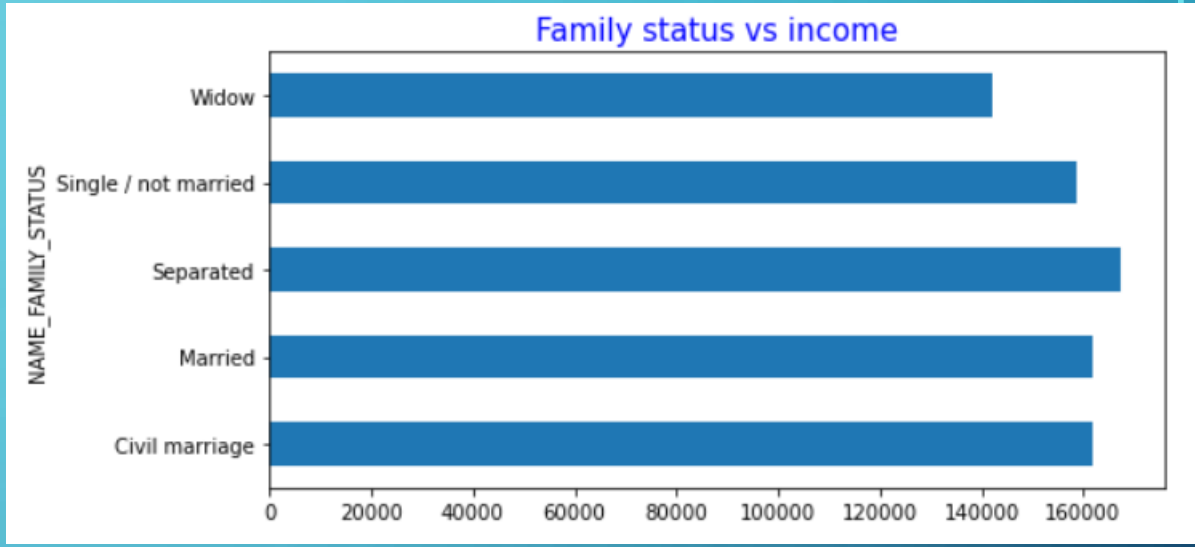
Bivariate analysis (Categorical vs Numerical)



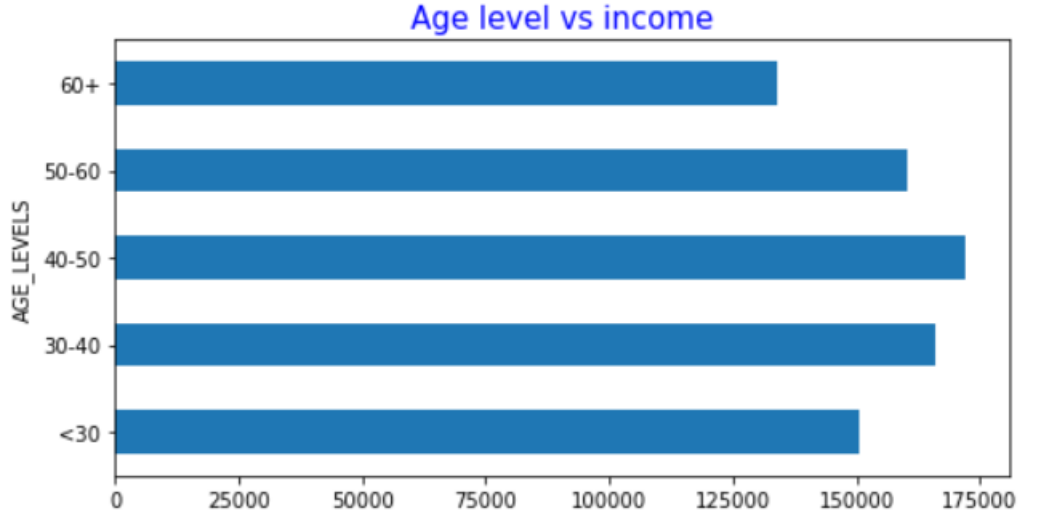
- Whisker plot on region rating vs income amount is observed here, although due to outliers it is difficult to understand the plot.
- Cities with higher ratings of 2 and 3 having their median almost at same level
- Whereas cities with rating 1 and salary near about 2.5 L has comparatively higher defaulters

Bivariate analysis (Categorical vs Numerical)

- From plot we can see that a person who is separated expected to have higher salary, and higher the salary lesser shall be the chances of defaulters.
- Similarly widowed are earning less and comparatively they have slightly higher chances of being defaulter than others
- Rest 3 cases seems to have similar pattern

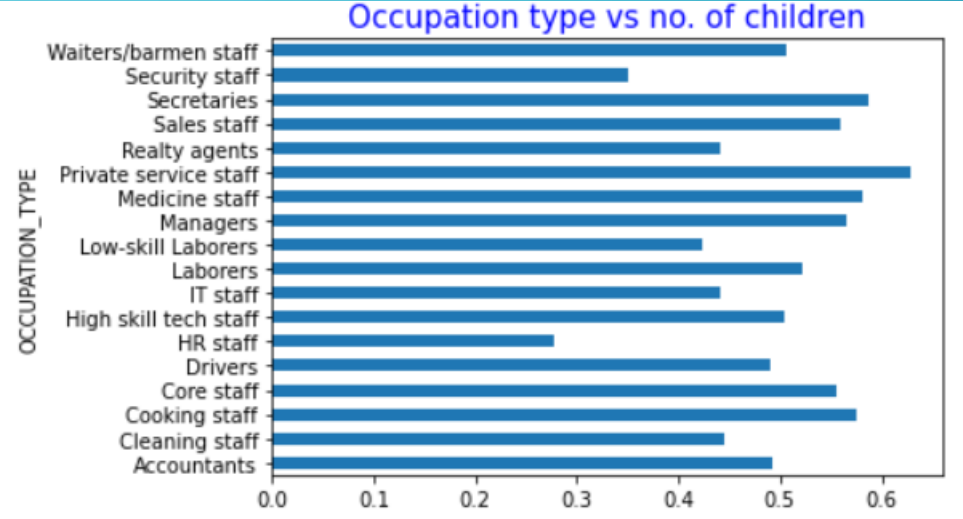


- People of age between 40-50 are having higher income, hence less likely to become a defaulter
- People of age 60+ are observed to have lesser salary and more likely to become a defaulter
- Also clients who have ages between 30-40 and 50-60 are following the same pattern

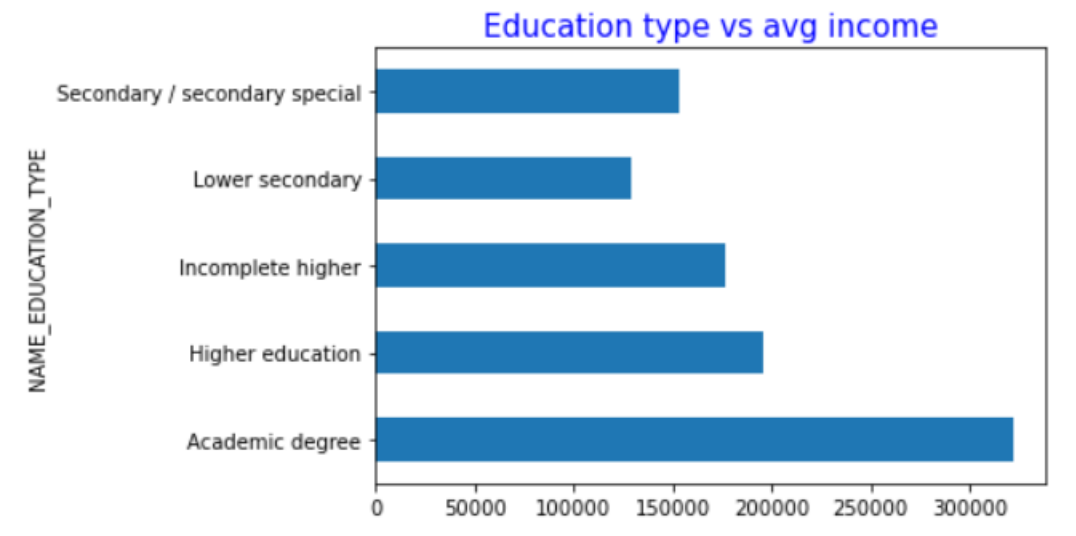


Bivariate analysis (Categorical vs Numerical)

- This can be observed that Private service staff are having higher no. of children hence more likely to have payment difficulties.
- A lower number around 3 is seen for HR staffs which is pretty low

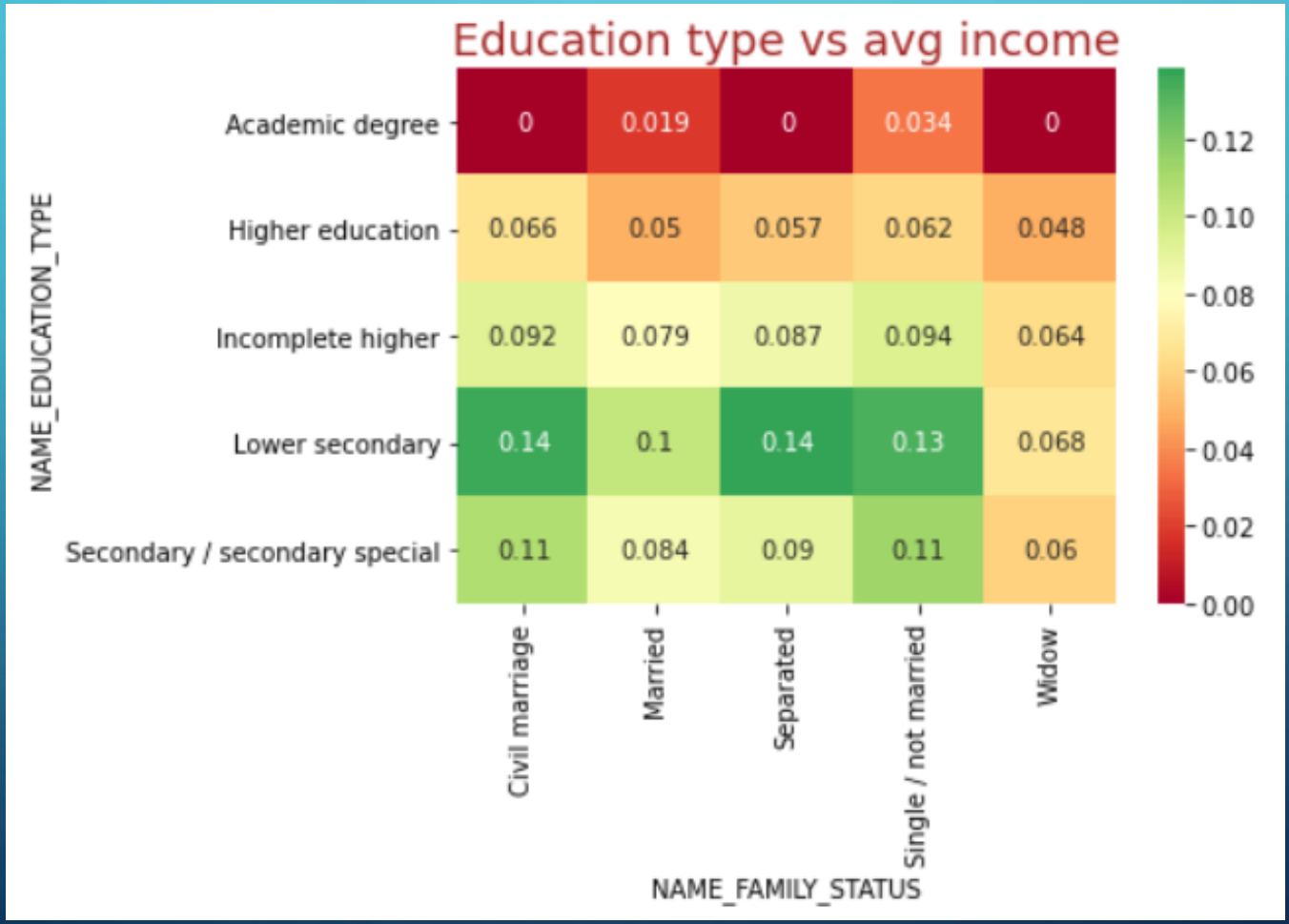


- This can be observed that clients with academic degree and has a high avg salary and has more no of defaulters.
- Although people with lower secondary has salary around 1.25L are having lesser repayment issues



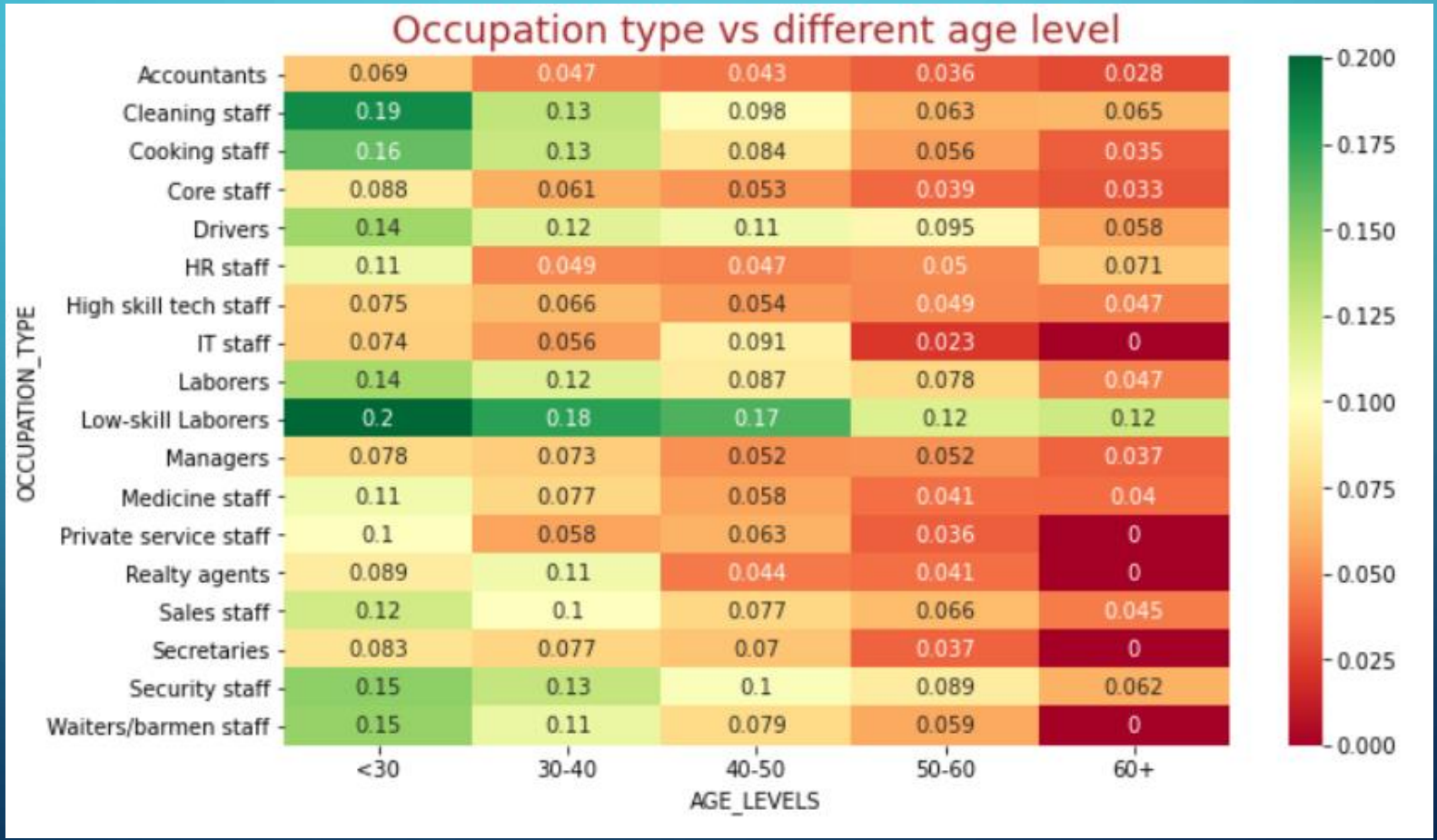
Multivariate analysis

- We can observe that the higher correlations will give us more no. of people to become defaulters,
- In this case we see that people with lower secondary education and with family status of civil marriage/separated are more likely to have repayment issues



Multivariate analysis

- Looks like the correlation is higher among the low-skill labours with ages below 30 thus these people are having higher chances to become defaulter
- Interesting point to observe that IT staffs who normally have less defaulter may give more defaulters if their age is in-between 40-50



Conclusion:

From the analysis so far conducted we can conclude the following

- Clients provided with cash loans are having more chances to become a defaulter
- Bank can make a through check on the age salary and occupation type to make a conclusion on whether to approve a loan.
- If a person belongs to a particular city with certain rating need to check on their total income to come to a conclusion to provide loans.
- The clients with previously approved loans are having more no of difficulties.
- A combination of different categorical correlations can be checked to come to a conclusion.