CREDIT EDA ASSIGNMENT

By Amaan Islamuddin Saifi

INTroDUCTIOn :

- Exploratory Data Analysis or EDA is a process of performing investigations to discover insights and patterns from the data set.
- It involves preparation of datasets by removing irregularities and also includes visualization techniques to make inferences.
- In this assignment we are going to use EDA for discovering insights from the datasets to understand what kind of variables are factors causing loan defaults.

Problem statement :

- The loan providing companies find it hard to give loans because of insufficient credit history of the customers.
- Banks or such companies have to take decision whether to accept a loan application or reject it.
- Such decisions have two risks:

1.Accepting the application even if the customer is likely to default ,then such a decision will cause financial loss to the company

2.Rejecting the application even if customer is likely to repay the loan,then such a decision will cause business loss to the company

Objective :

- We aim to identify those customers who are likely to repay the loan through EDA process.
- Basically , understanding the driving factors behind loan default is our objective.

ASsumPTIons :

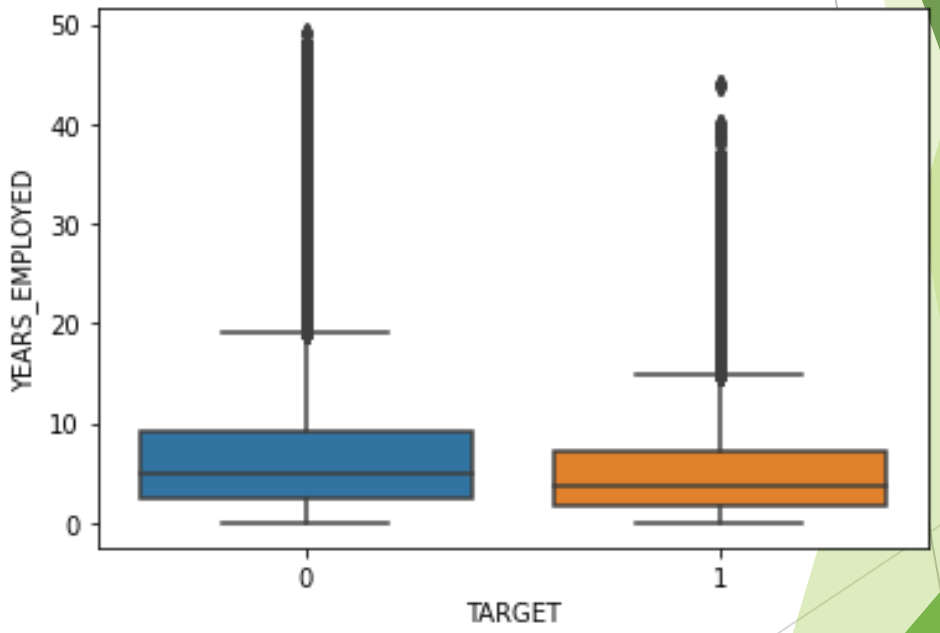
- Some columns had XAP and XNA values . I assumed this to be null values so I had to replace it by null .
- 1 column had some -1 values . Again ,I assumed it to be null.
- In the dataset there were outliers but I didn’t remove them all.I kept some outliers but some inexplicable outliers had to be dropped.

APPROACH :

- Understanding the domain and columns.
- Loading the datasets in dataframes.
- Checking shape and datatypes.
- Missing value treatment : Dropping the missing values which were interfering in analysis and deleting missing values in each column if they were more than 40 %.
- Outlier analysis : Dropping some inexplicable outlers while keeping some outliers.
- Merging the two dataframes using left join.
- Univariate Analysis and Bivariate Analysis.

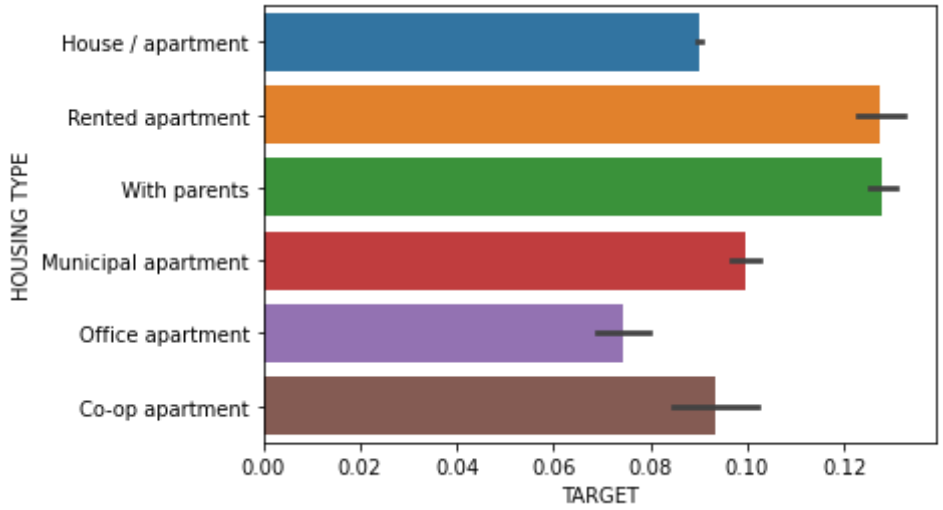
YEARS_EMPLOYED :

In the boxplot we can observe that years_employed median is more for clients having no difficulties in payment than clients having payment difficulties.The 75 percentile of years employed for clients with payment difficulties is also much less than 75 percentile of years employed for client with no payment difficulties.



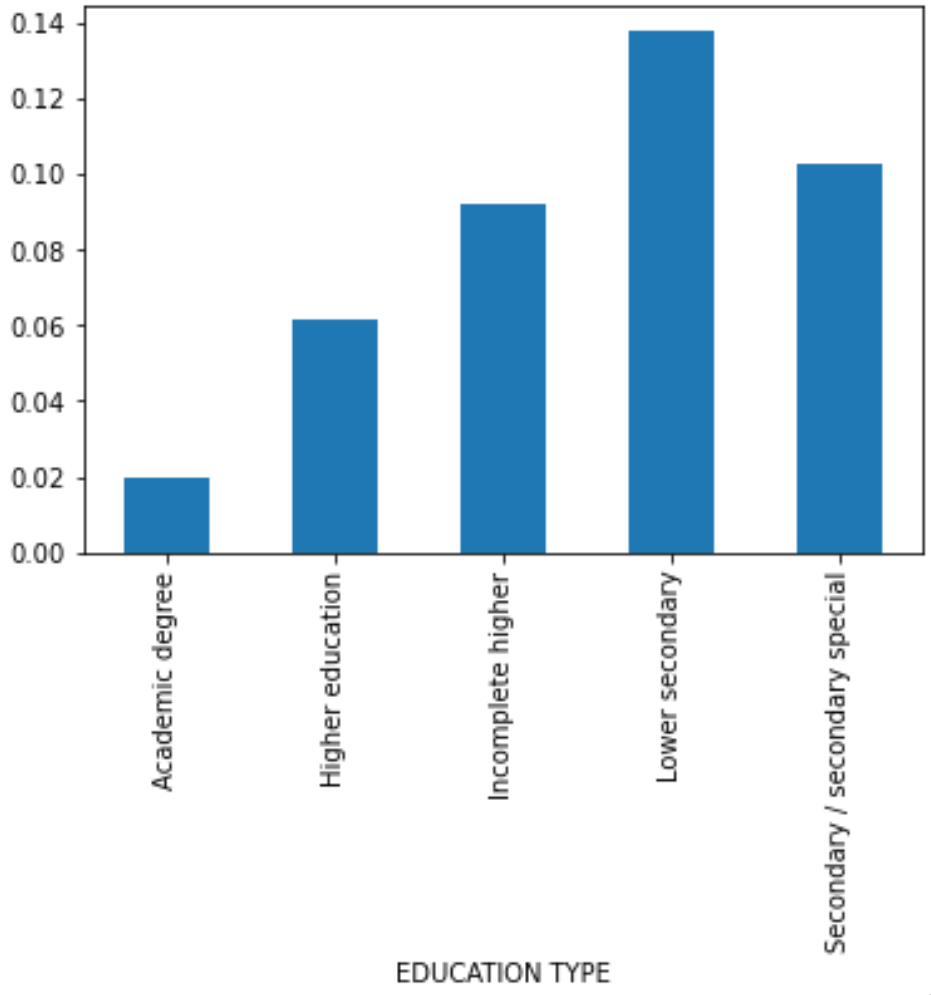
HOUSING TYPE :

We can infer that customers living in a rented apartment and cutomers living with parents had more tendency to default than customers living in their own apartment, or customers living in an office apartment.



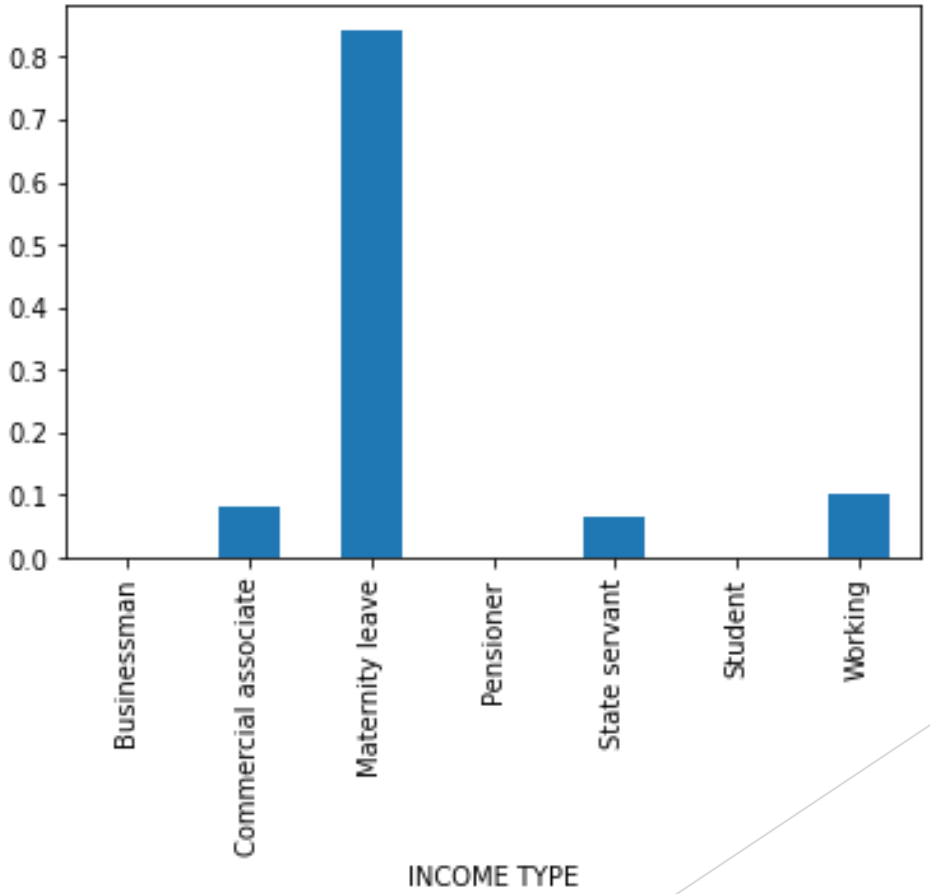
Education type

Customers having lower secondary as their education background had more tendency to default than customers having academic degree.



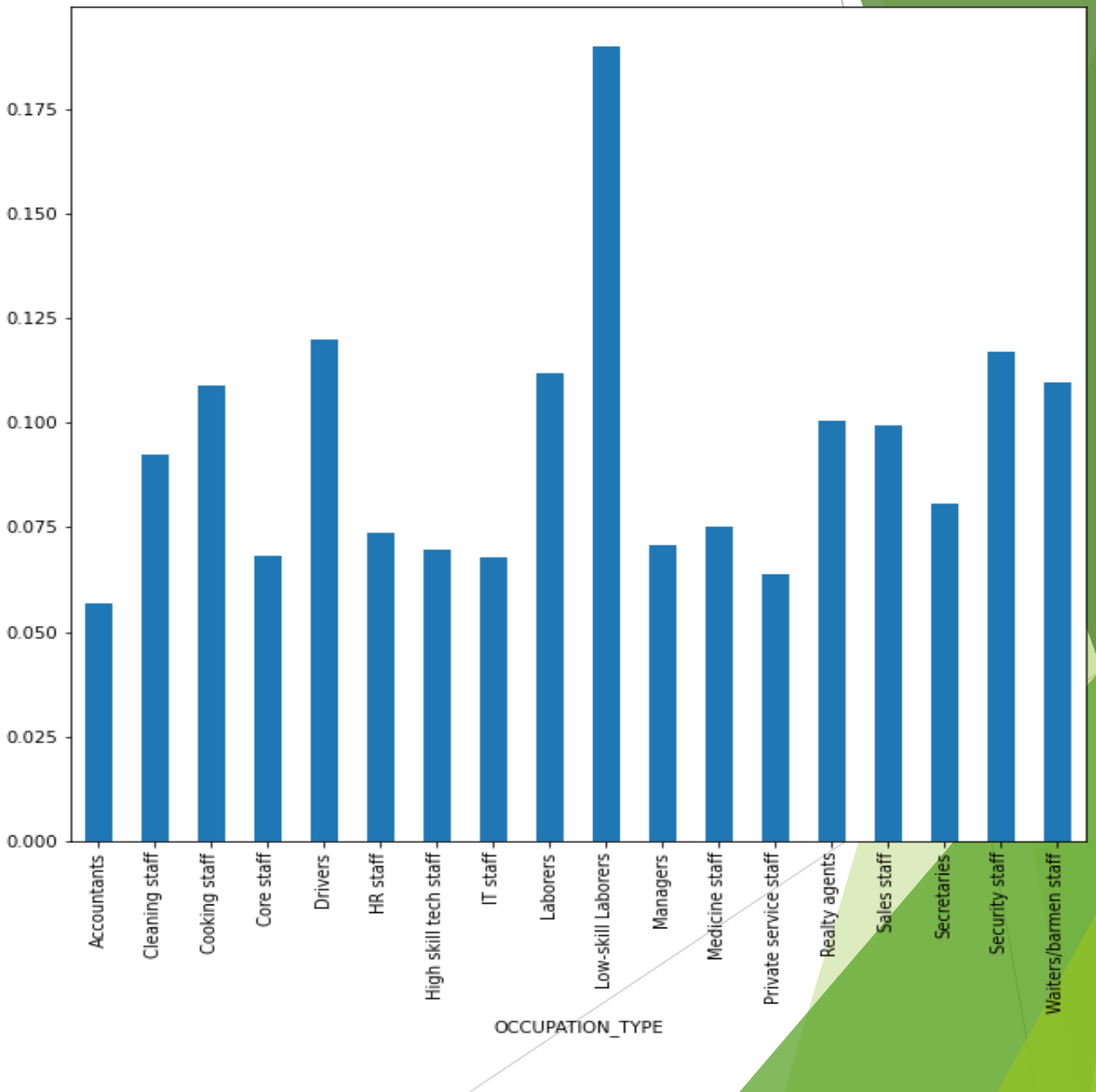
INCOME TYPE:

Customers having their income type maternity leave had more tendency to default than customers who had their income type as state servant and working.



OCCUPATION TYPE

Customers having occupation type as low skill laborers had more tendency to default on their loan than customers who were accountant , managers or were in IT staff , Core staff , Private Service staff.



Conclusion :

- Through EDA we found useful patterns and insights
- Because of such insights the loan giving company can identify customers who are capable of repaying their loans.
- On the basis of these patterns the company can reject the loan or reduce the loan amount or increase the interest rate on loan.