EDA On **Bank** Loan

Submitted By Harshita Mundhe

PROJECT DESCRIPTION

In this project we have given a dataset and using this dataset we have to perform EDA. Exploratory Data Analysis is a technique which is used to identify the various patterns from the data. It allows us to analyze the data before coming to any assumption. It ensures that the results produced are valid and applicable to business outcomes and goals.

The dataset which is given in this project is regarding the bank loan. The problem is that the loan providing company is finding difficulty in giving loan to customer due to their insufficient credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. So we have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

Basically we have to use EDA technique to identify if a client has difficulty paying their dues so that no loan is approved which could default and no clients who are capable of paying the loan are rejected and identifying variables which are strong indicators of default.

APPROACH

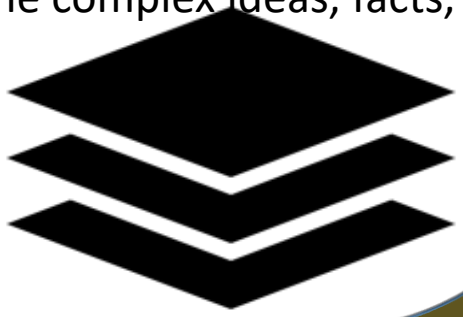
For implementing this project I have used Python and Jupyter Notebook. I performed various operation on dataset to find out different patterns and trends like data understanding, data cleaning, finding out outliers, finding data imbalance ratio and many more so that it will ensure that the applicants capable of repaying the loan are not rejected.

TECH-STACK USED

- **Python:** I have used python for EDA as Python is a popular programming language in scientific computing, because it has many data-oriented feature packages that can speed up and simplify data processing, thus saving time. Also it provides users with a plethora of different visualization options.

- **Jupyter Notebook:** I have used Jupyter Notebook as it provides a feature-rich, robust, and user-friendly environment using multiple installation methods. Jupyter allows us to view the results of the code in-line without the dependency of other parts of the code

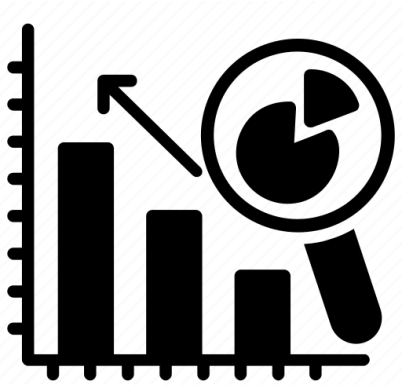
- **PowerPoint Presentation:** I have used Microsoft PowerPoint 2019 MSO (Version 2212 Build 16.0.15928.20196) 64-bit to create a report as it allow us to present the complex ideas, facts, or figures into easily digestible visuals.



INSIGHTS

- **Reading and Understanding the data:-**

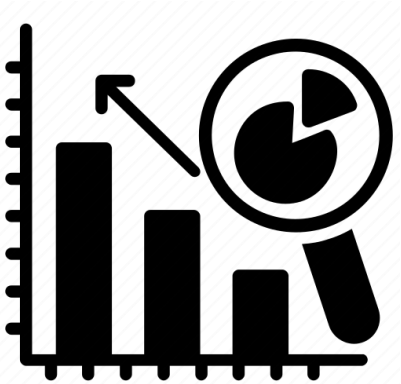
- To perform EDA on dataset we have to read the data and understand it.
- For reading the data I have used Python's Pandas library.
- For Understanding data and variables present in it I have used methods like head(), info(), shape, describe, etc.



INSIGHTS

- **Removing/Imputing Missing Values:-**

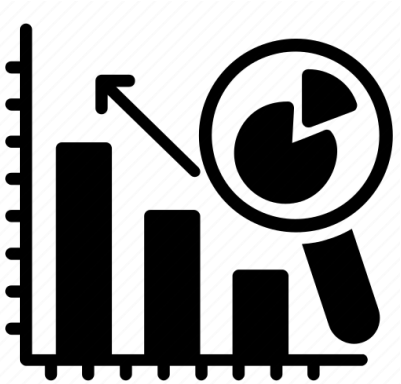
- Data with missing value more than 40% has been removed from both the datasets.
- Columns which does not have any use in analysis has been removed by finding out their description.
- Some columns has been removed by drawing correlation between them.
- Remaining columns having missing value has been imputed using mean, median, mode as per the need and requirement.
- Converted some continuous value into categorical data such as using Days_Birth column AGE column is derived.



INSIGHTS

- **Finding outliers:-**

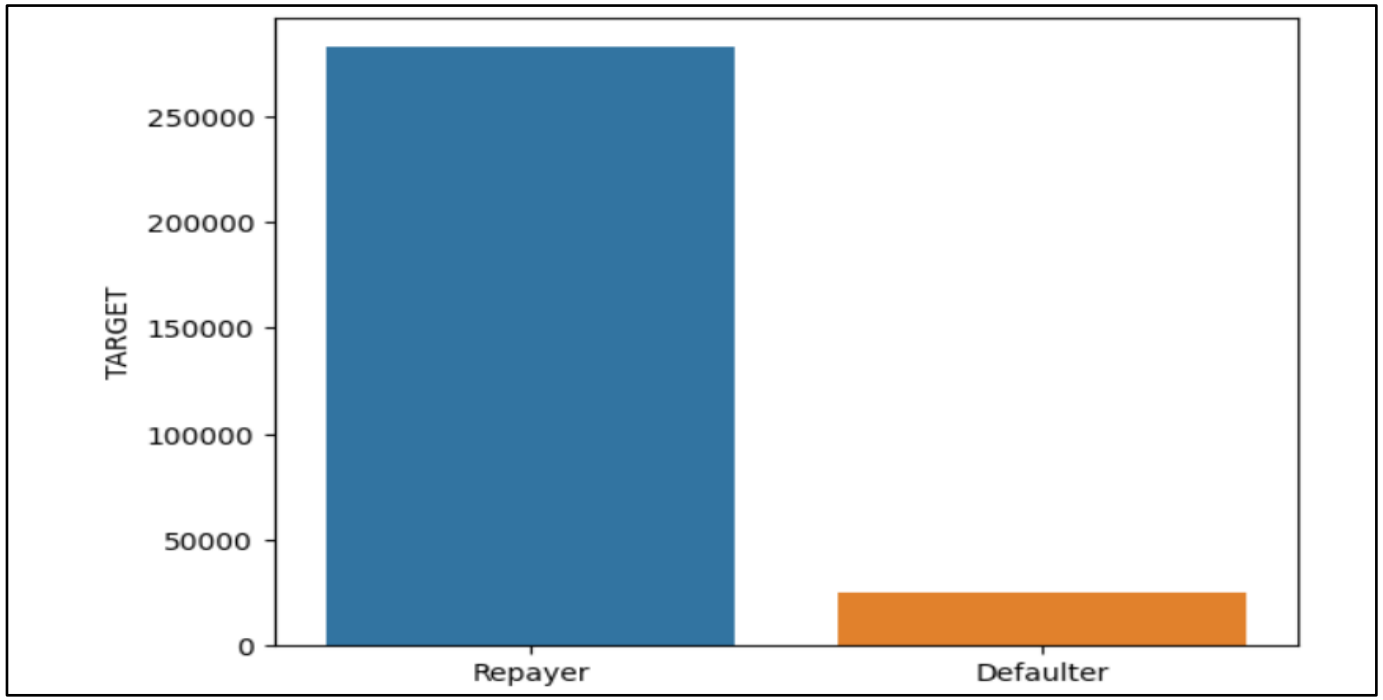
- For finding outliers boxplot has been used which is available under seaborn library.
- AMT_INCOME_TOTAL has outlier near 12 crore.
- AMT_CREDIT has many outliers lying above 20 Lakh.
- AMT_ANNUITY has many outliers lying above 20 Lakh.
- DAYS_BIRTH does not have any outlier meaning that it does not contain any incorrect value and data is reliable.
- DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.

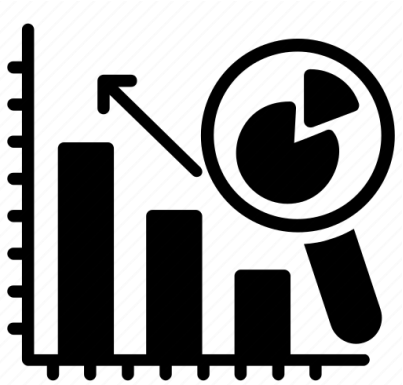


INSIGHTS

- **Data Imbalance ratio:-**

➢ In the given dataset imbalance ratio is 92:8.



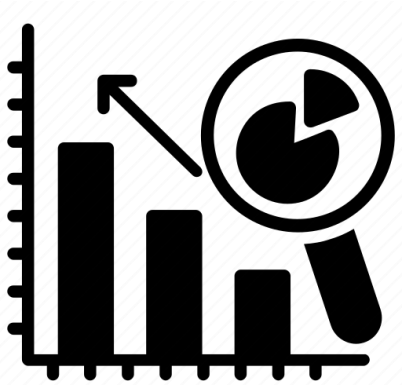


INSIGHTS

- **Correlated Variable:-**

- As data is imbalanced it is divided into two parts based on defaulter and non defaulter.
- Top 10 correlated variable list of defaulter is as follows:

	var1	var2	corr
1096	DAYS_EMPLOYED_YEARS	DAYS_EMPLOYED	1.000000
1129	AGE	DAYS_BIRTH	0.999691
803	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269
173	AMT_GOODS_PRICE	AMT_CREDIT	0.982783
454	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637
375	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
838	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868994
594	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885
699	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540
174	AMT_GOODS_PRICE	AMT_ANNUITY	0.752295

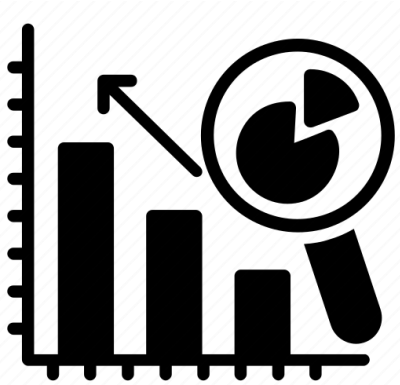


INSIGHTS

- **Correlated Variable:-**

➢ Top 10 correlated variable list of non defaulter/repayers is as follows:

	var1	var2	corr
1096	DAYS_EMPLOYED_YEARS	DAYS_EMPLOYED	1.000000
1129	AGE	DAYS_BIRTH	0.999711
803	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508
173	AMT_GOODS_PRICE	AMT_CREDIT	0.987022
454	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149
375	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
594	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861
838	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859332
699	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830381
174	AMT_GOODS_PRICE	AMT_ANNUITY	0.776421

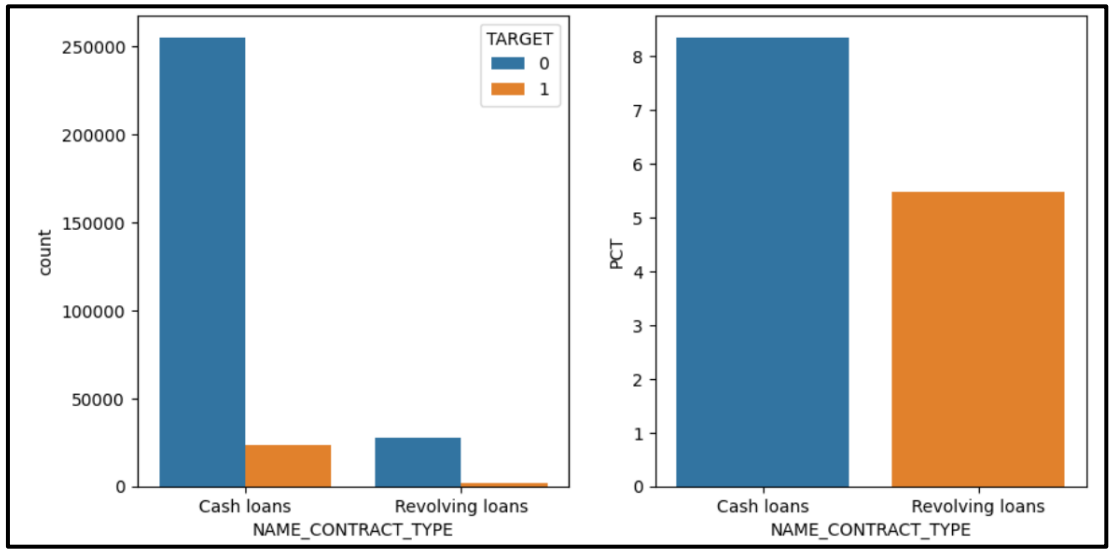


INSIGHTS

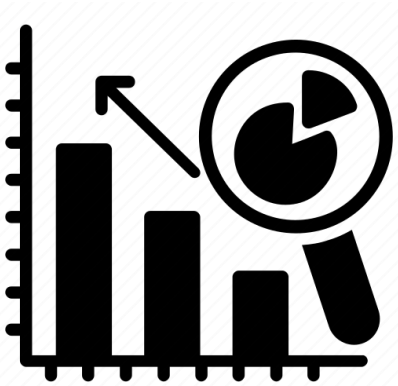
- **Data Analysis:-**

➢ Univariant Analysis on categorical values:

▪ Contract Type:

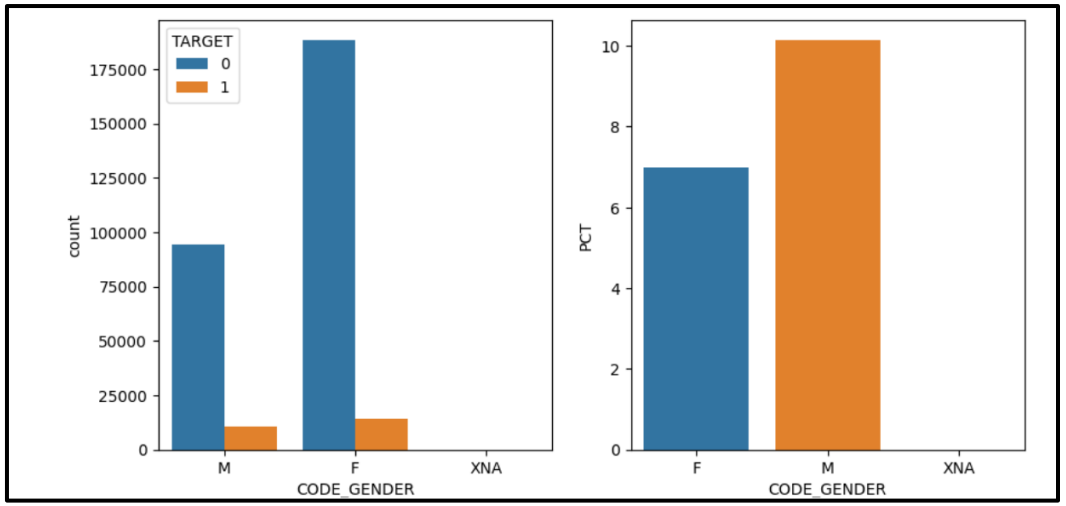


- Most of the customers have taken Cash loans.
- Customers who have taken cash loans are less likely to default.



INSIGHTS

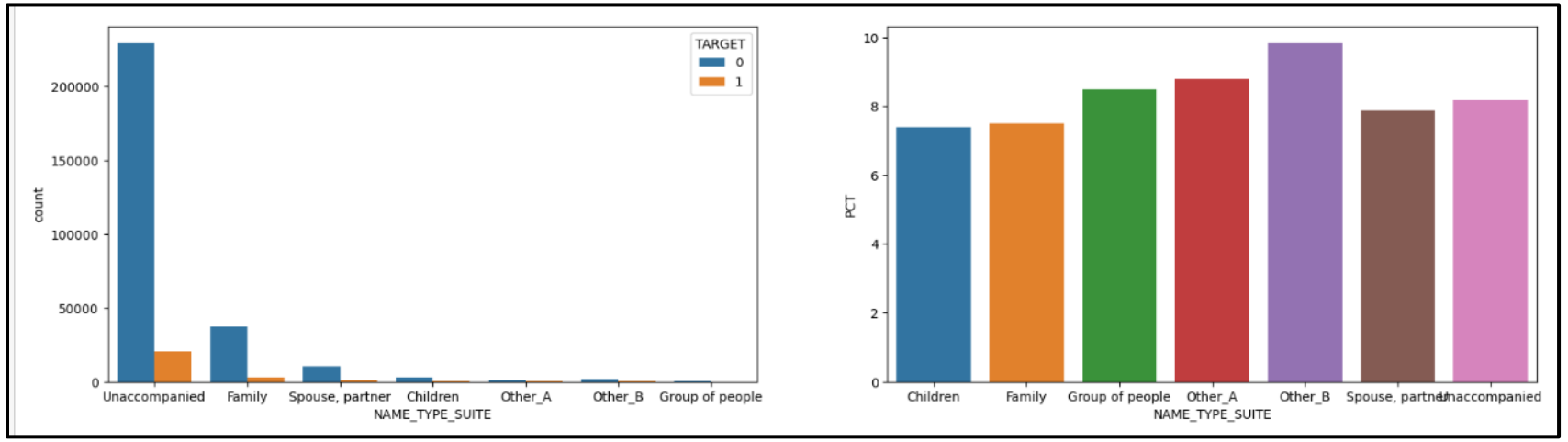
- **Data Analysis:-**
- Gender:



- Most of the loans are taken by female.
- Females are safer customer as they have less default rate as compared to male..

INSIGHTS

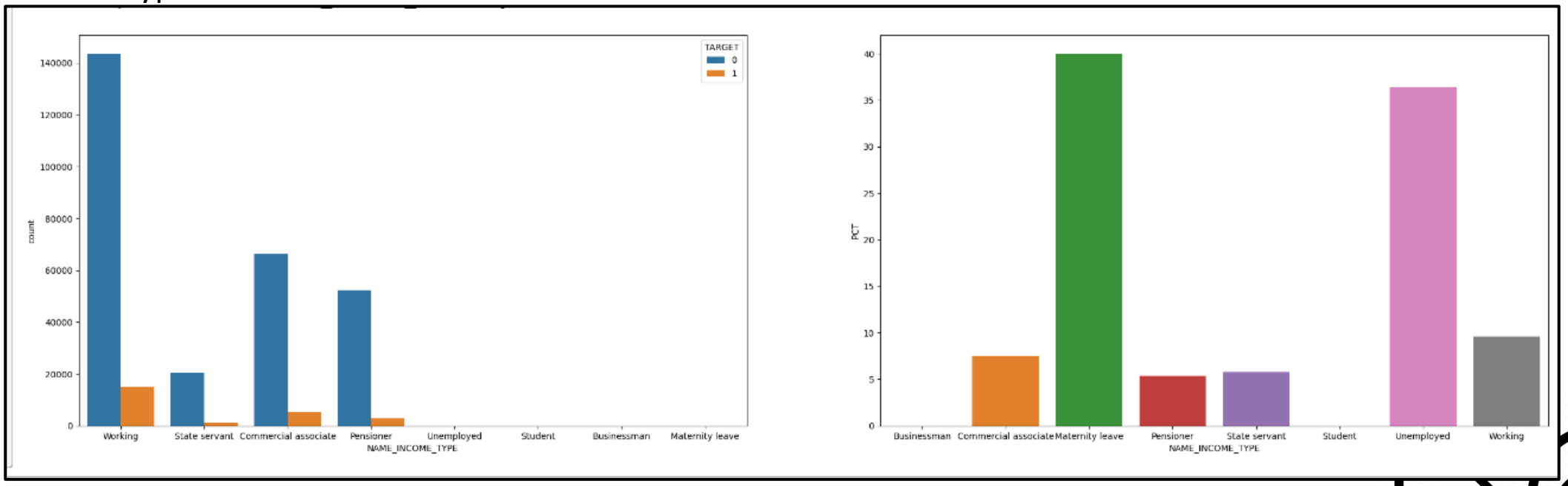
- **Data Analysis:-**
  - Type Suite:



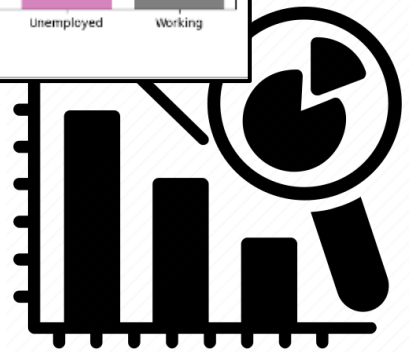
- Most of the loans are taken by Unaccompanied.
- Unaccompanied are safer customer as they have less default rate as compared to others.

INSIGHTS

- **Data Analysis:-**
- Income Type:

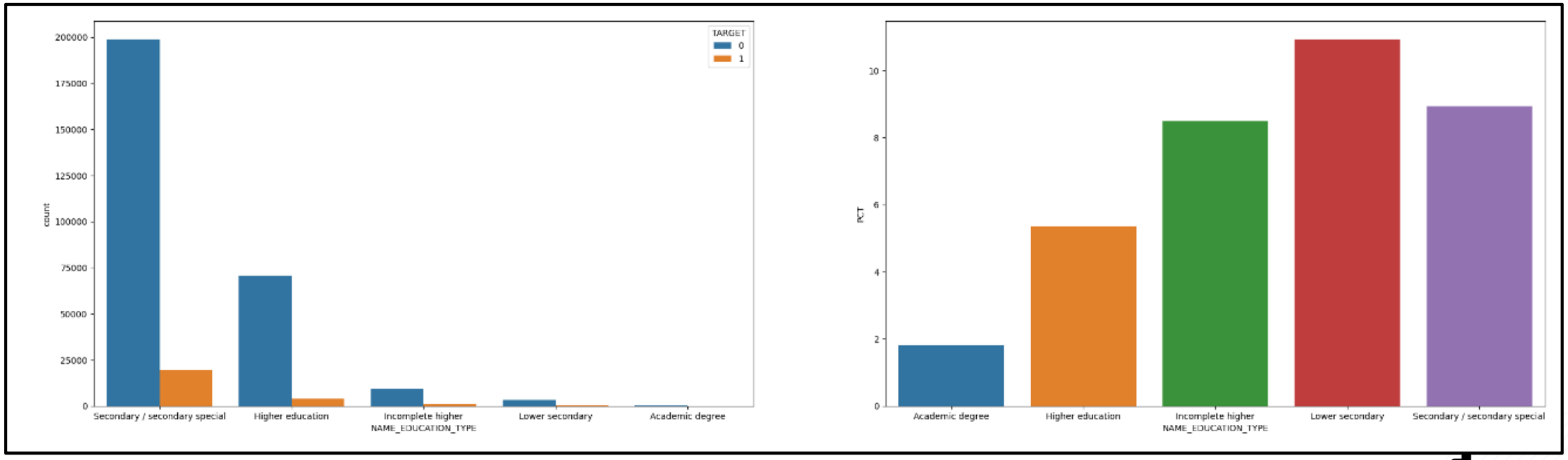


- Working, commercial associates and pensioners are safest segment

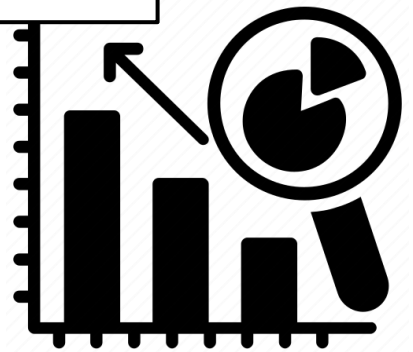


INSIGHTS

- **Data Analysis:-**
- Education Type:

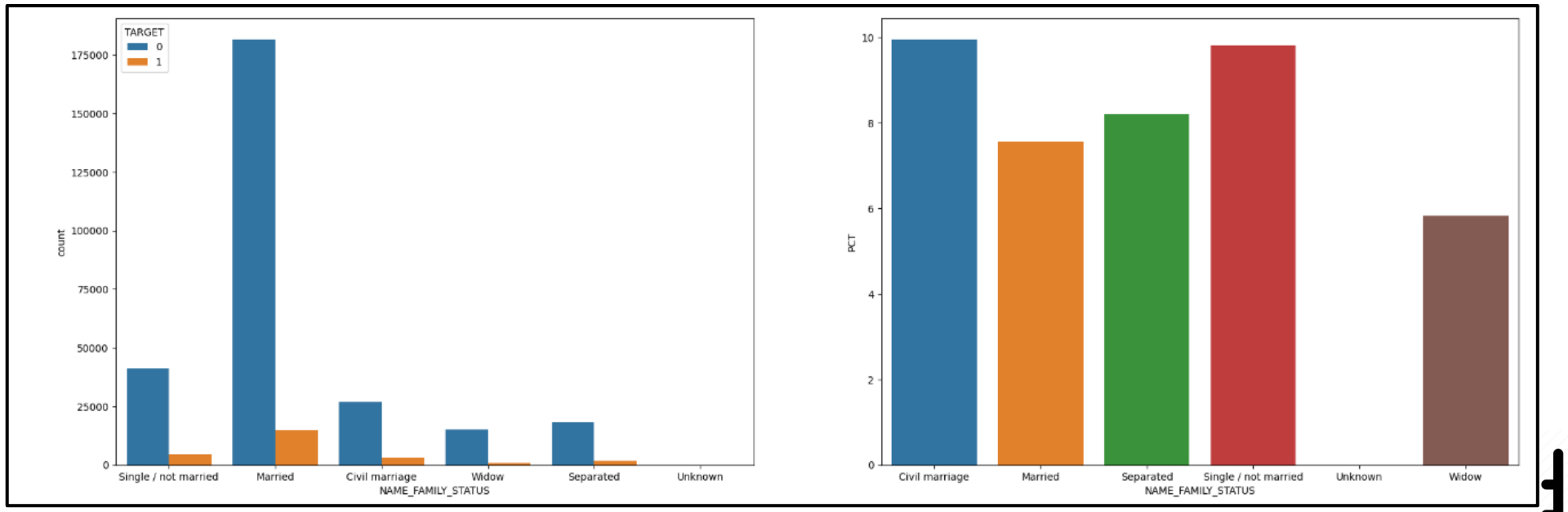


- Higher education is the safest segment to give loan as it has less default rate.

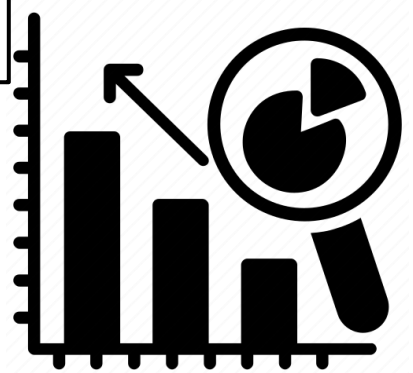


INSIGHTS

- **Data Analysis:-**
- Family Status:

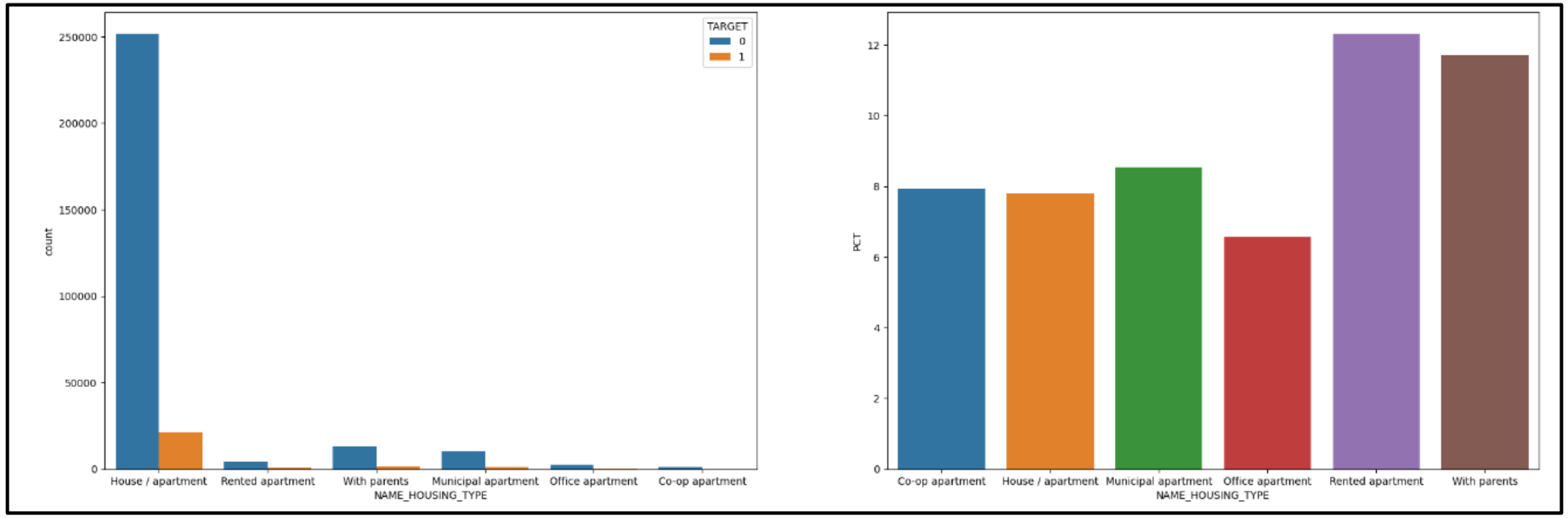


- Married people are safest to give loans as they have less default rate.

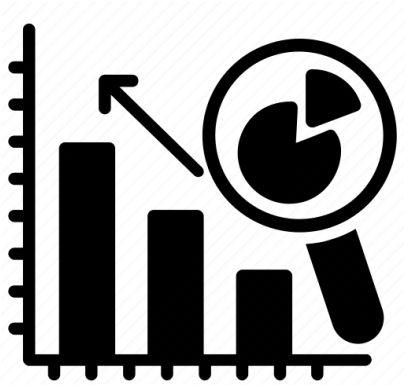


INSIGHTS

- **Data Analysis:-**
  - Housing Type:

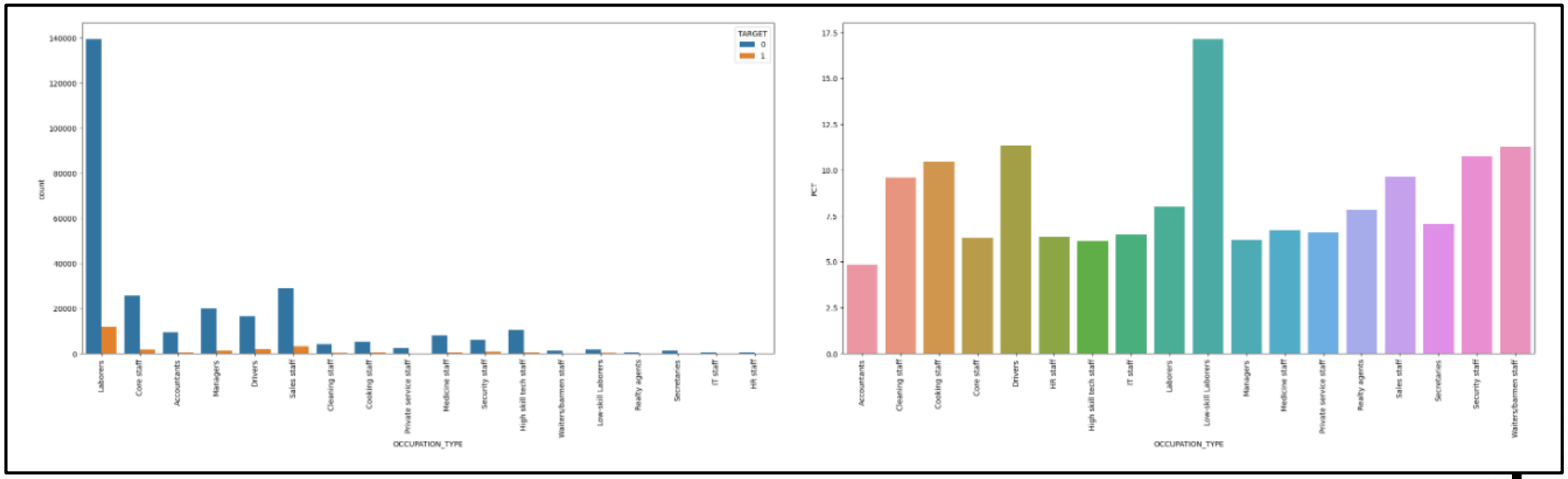


- Customers having house/apartment are safest segment to grant loans

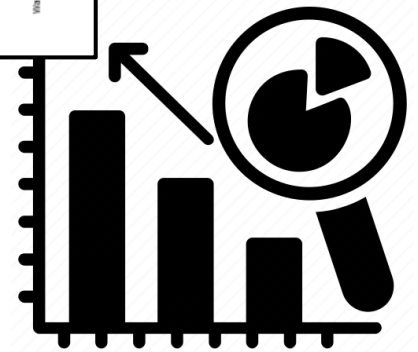


INSIGHTS

- **Data Analysis:-**
  - Occupation Type:

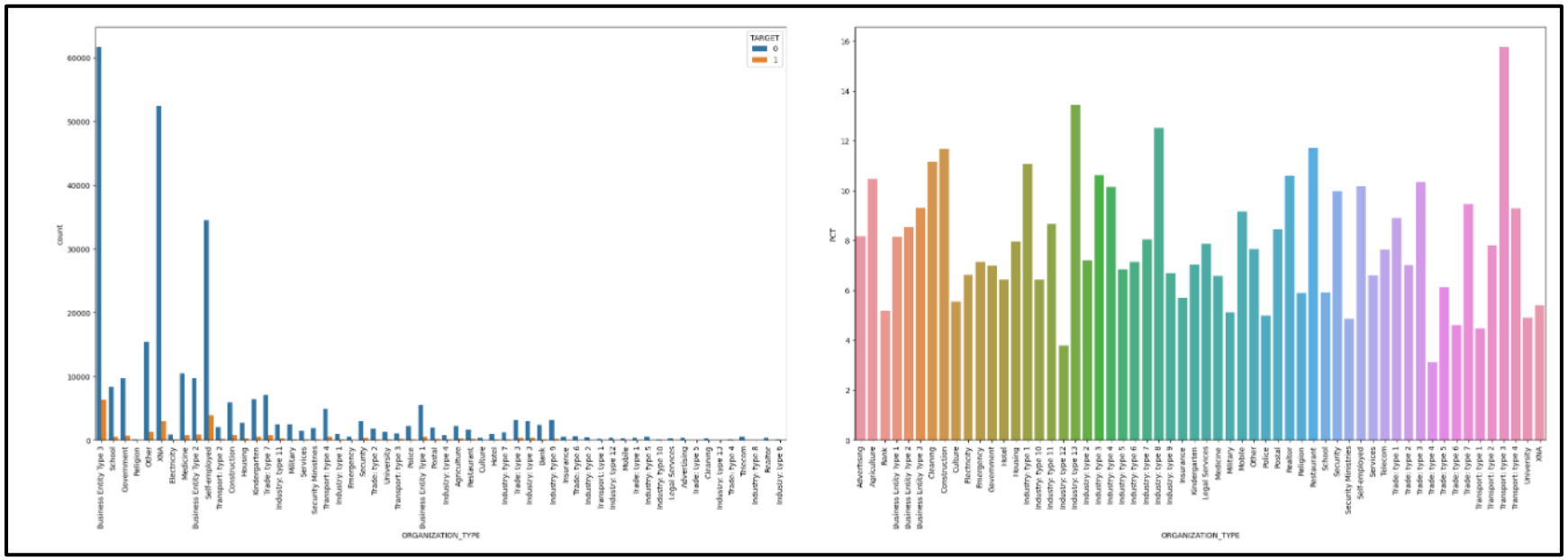


- Low-Skill Laborers and drivers have the highest default rate.
- Core staff, Managers and Laborers are safest segment to give loans.

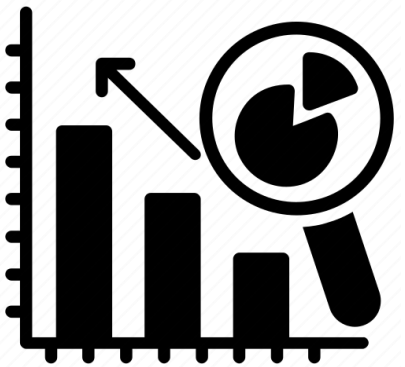


INSIGHTS

- **Data Analysis:-**
- Organization Type:



- Transport type 3 has highest default rate.
- Others, Business Type 3, Self Employed are safest with highest rate around 10%.

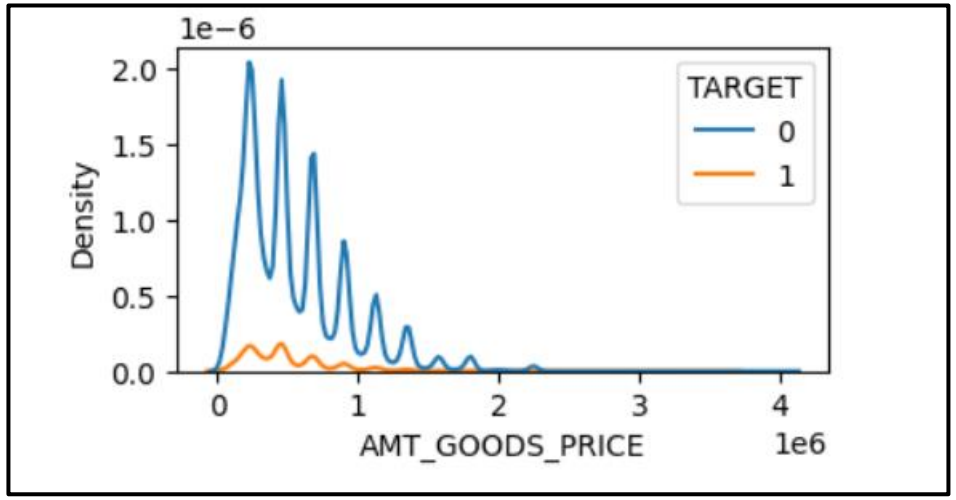


INSIGHTS

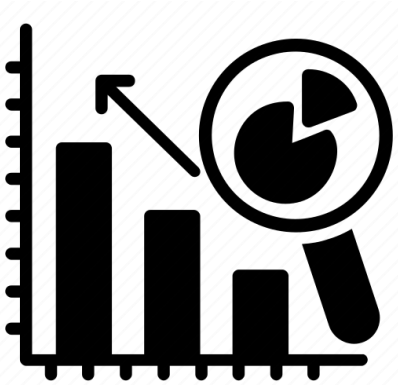
- **Data Analysis:-**

➢ Univariant Analysis on Numeric values:

- Goods Price:



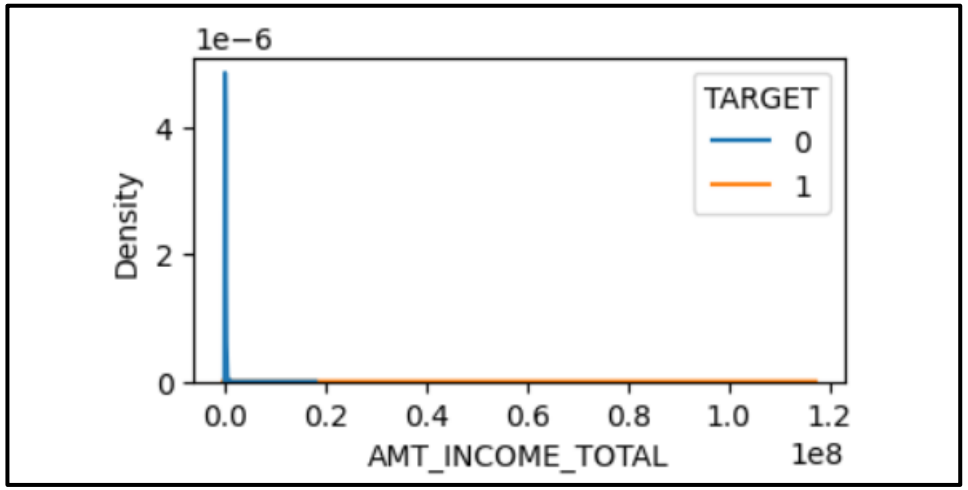
- Most of the loans given for the goods price ranging between 0 to 1 million.



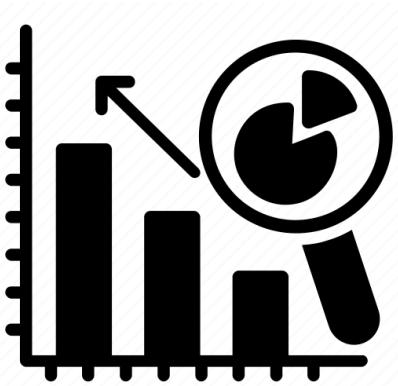
INSIGHTS

- **Data Analysis:-**

- Income:



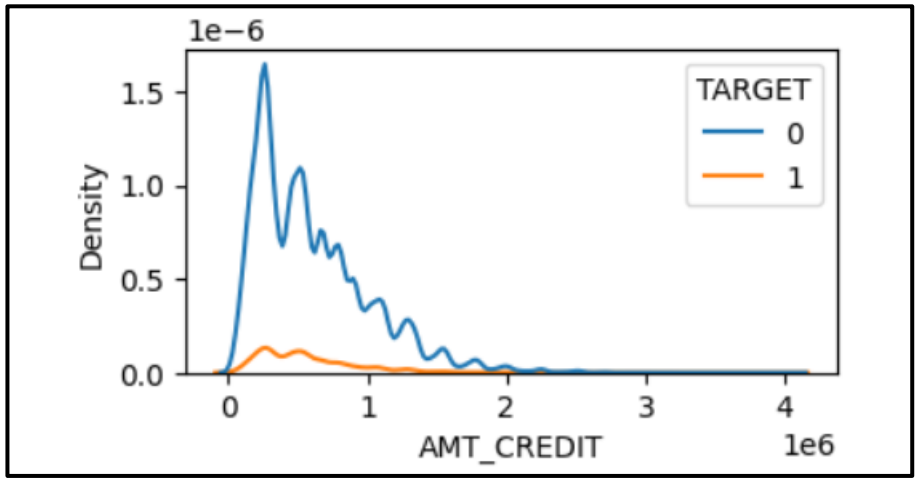
- Most of the customers have income between 0 to 1 ml.



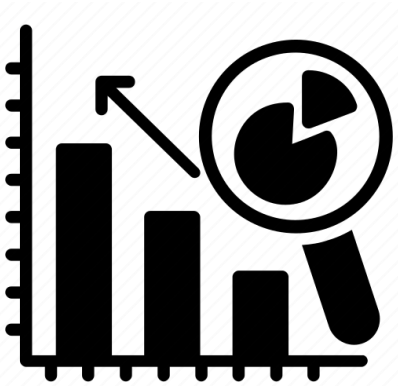
INSIGHTS

- **Data Analysis:-**

- Credit Amount:



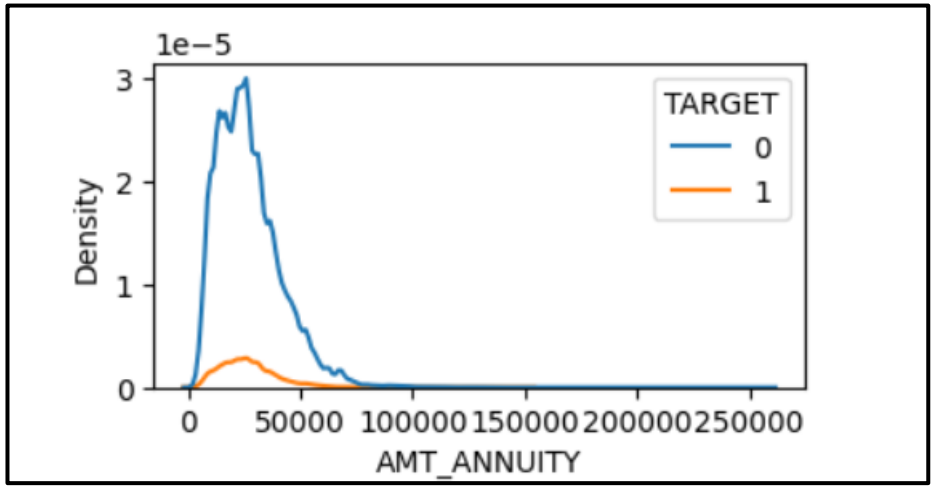
- Most of the loans were given for the credit amount of 0 to 1 million.



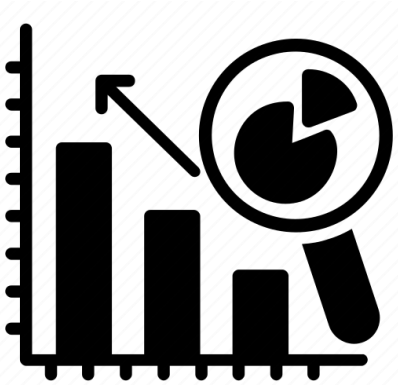
INSIGHTS

- **Data Analysis:-**

- Annuity Amount:



- Most of the customers are paying annuity of 0 to 50k.

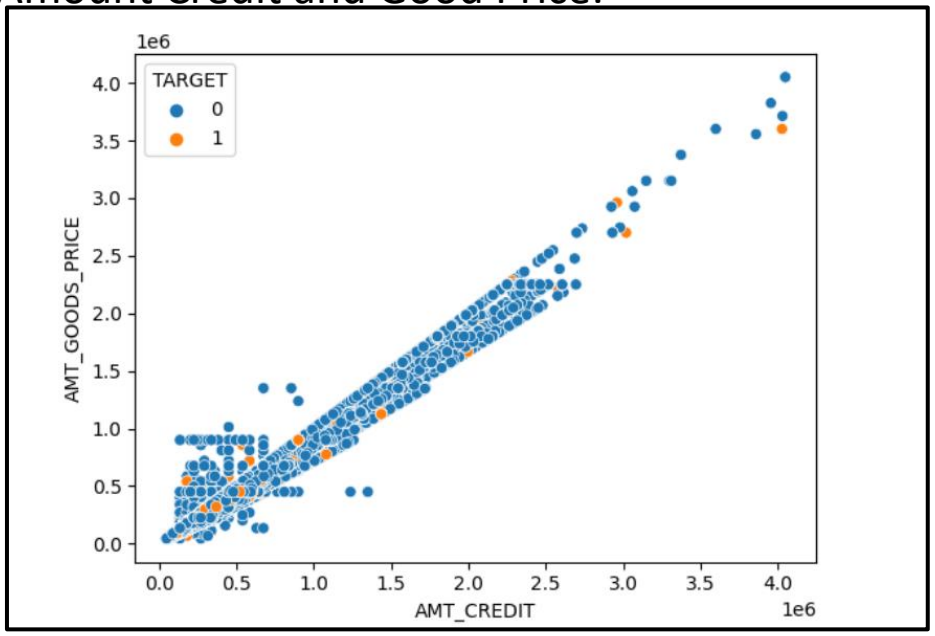


INSIGHTS

- **Data Analysis:-**

➢ Bivariant Analysis:

▪ Amount Credit and Good Price:

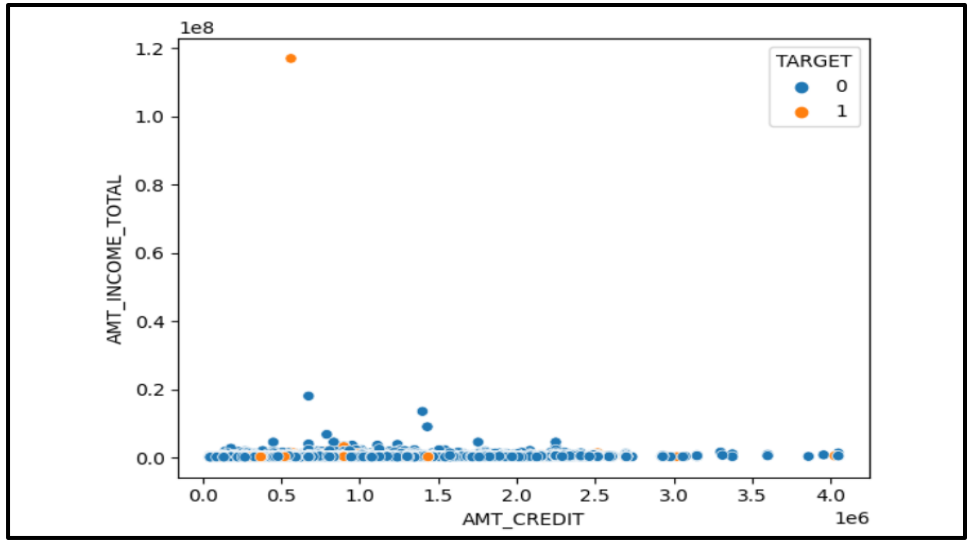


- AMT_CREDIT and AMT_GOODS_PRICE are linearly correlated, if the credit amount increases the defaulters are decreasing.

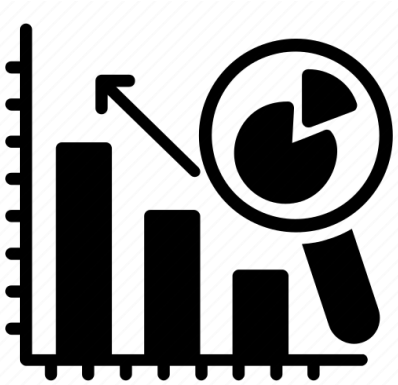
INSIGHTS

- **Data Analysis:-**

- Credit Amount and Income:



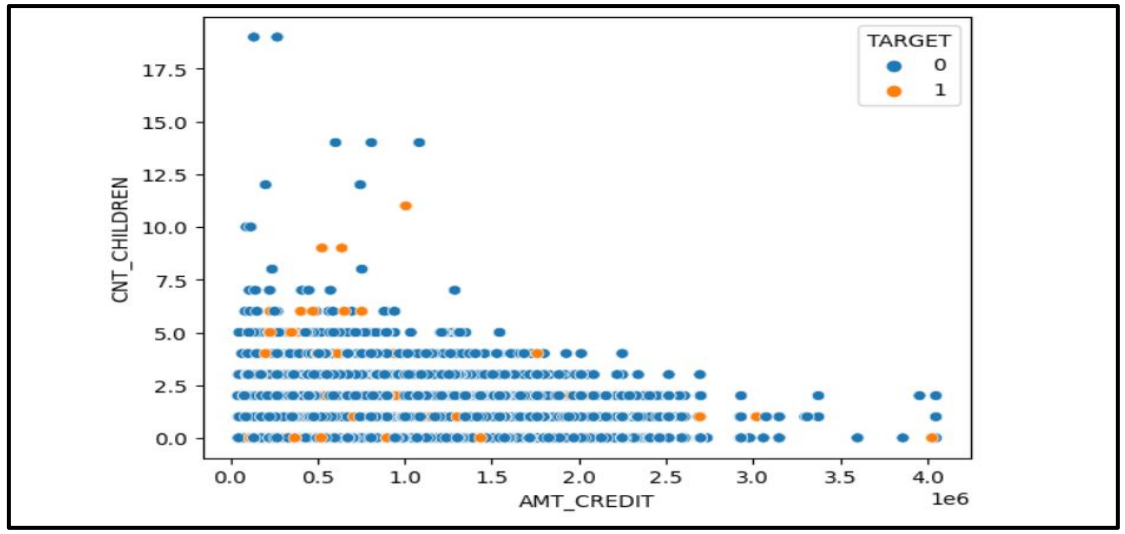
- Customers having income less than or equals to 1 m, are more like to take loans out of which who are taking loans less than 1.5 million, could turn out to be defaulters.



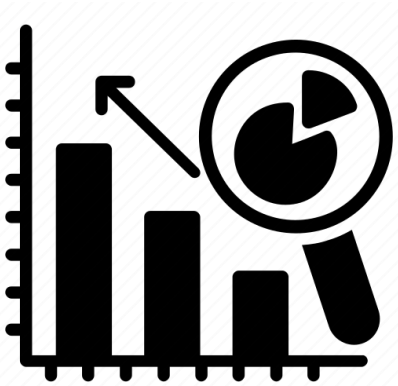
INSIGHTS

- **Data Analysis:-**

- Credit Amount and Income:



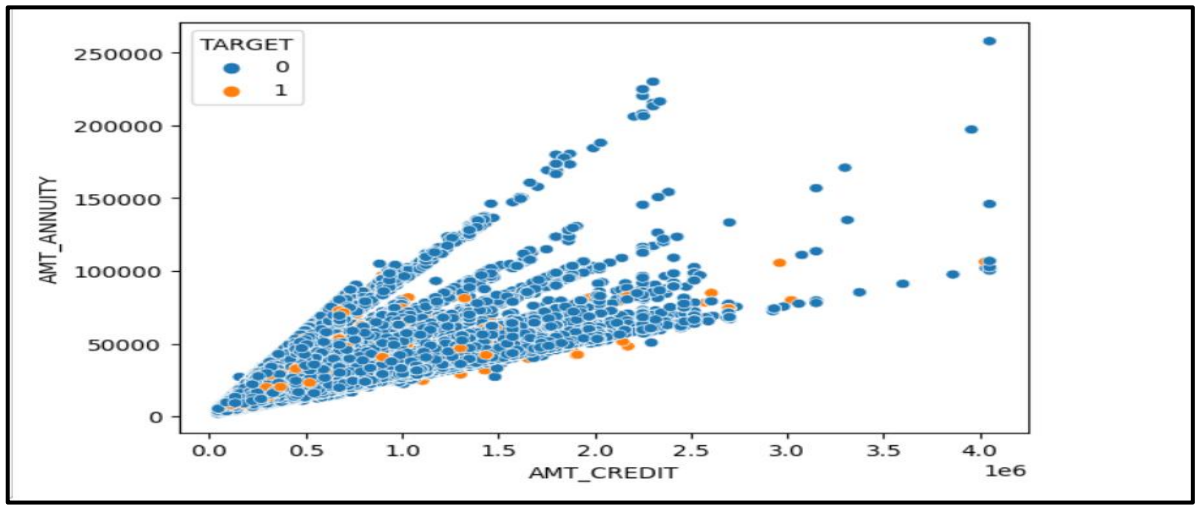
- Customers having 1 or less than 5 children can be considered to give loan.



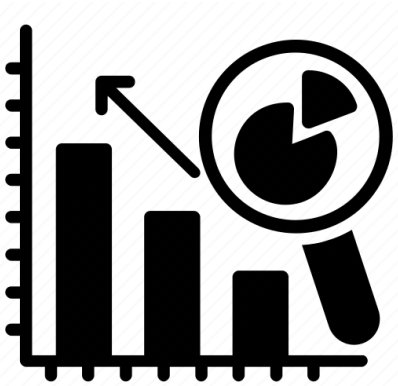
INSIGHTS

- **Data Analysis:-**

- Credit Amount and Annuity Amount:



- Customers who can pay annuity of 100k can be considered to give loans.

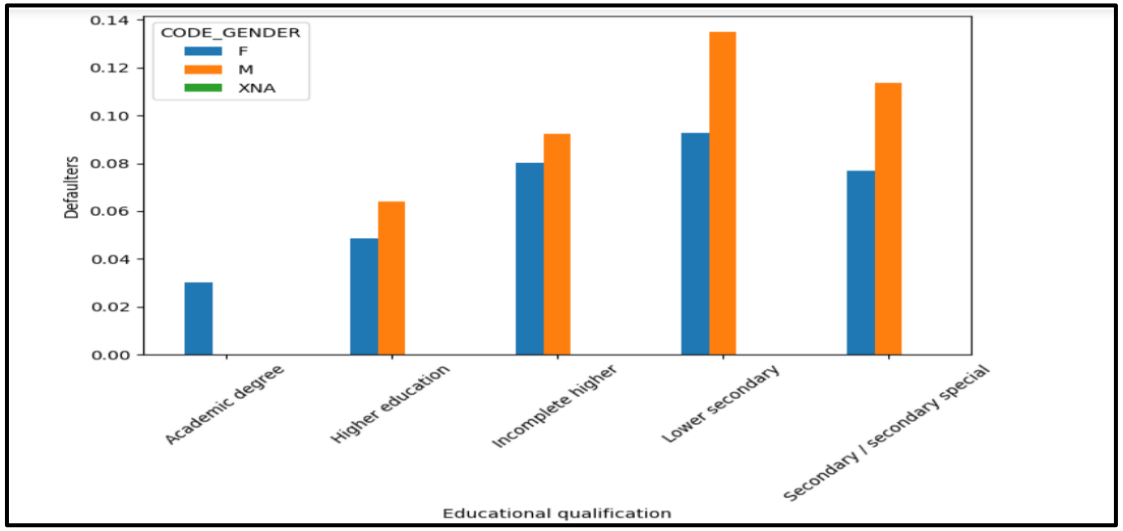


INSIGHTS

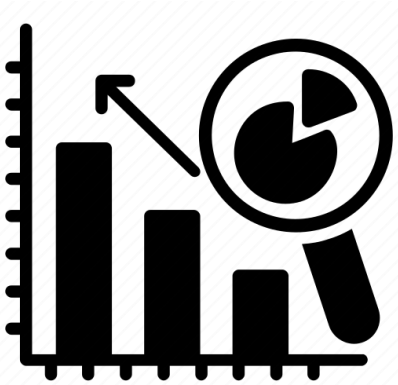
- **Data Analysis:-**

➢ Segmented Analysis:

▪ Education Type & Gender:

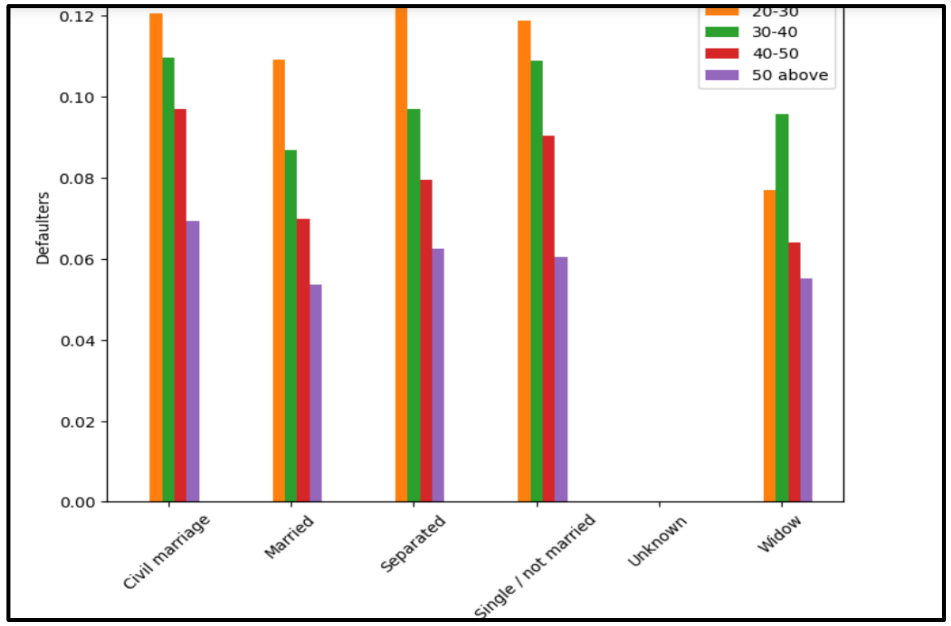


- Higher educated people are less defaulted and lower secondary educated people are more.

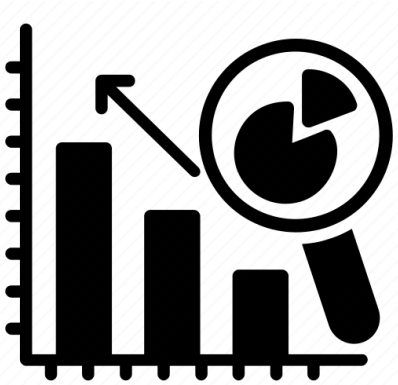


INSIGHTS

- **Data Analysis:-**
- Family Status & Age Group:

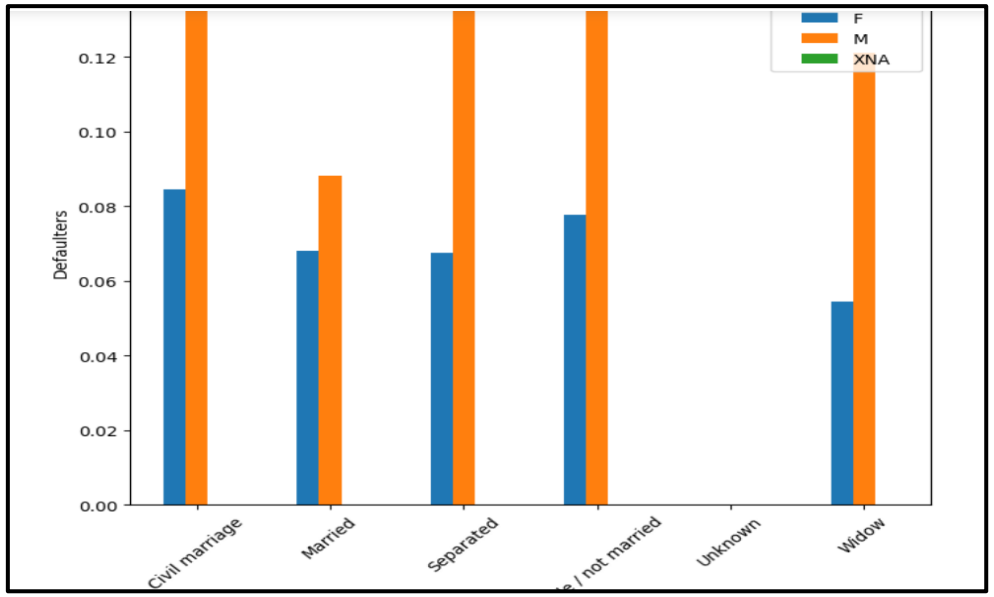


- Senior citizen can be considered to give loans as irrespective of family status they are less likely to be defaulted.
- Young people are more likely to be defaulted in all family status.

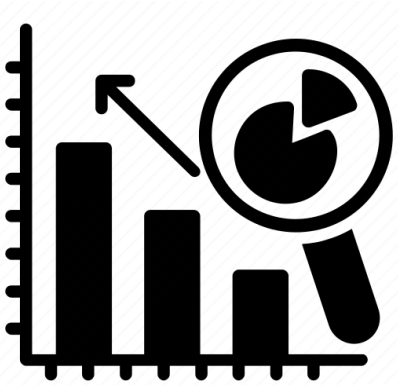


INSIGHTS

- **Data Analysis:-**
  - Family Status & Gender:

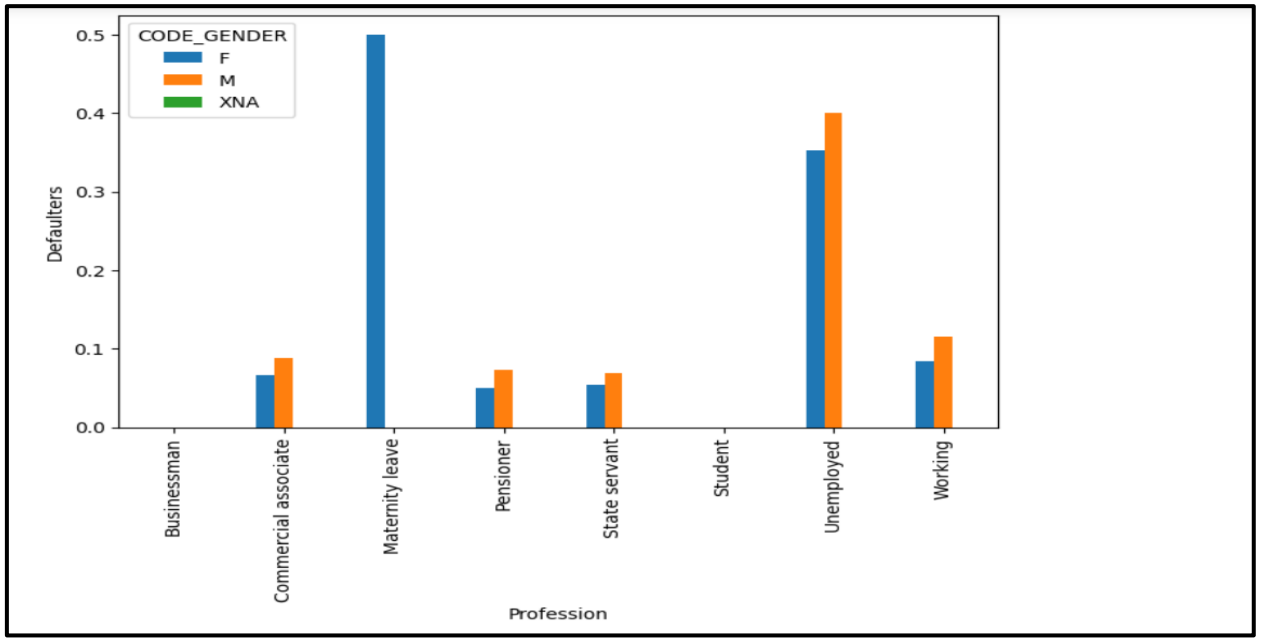


- Males are more like to be defaulted than females.
- It is not recommended to grant loan for singles, separated and civil marriage young men.



INSIGHTS

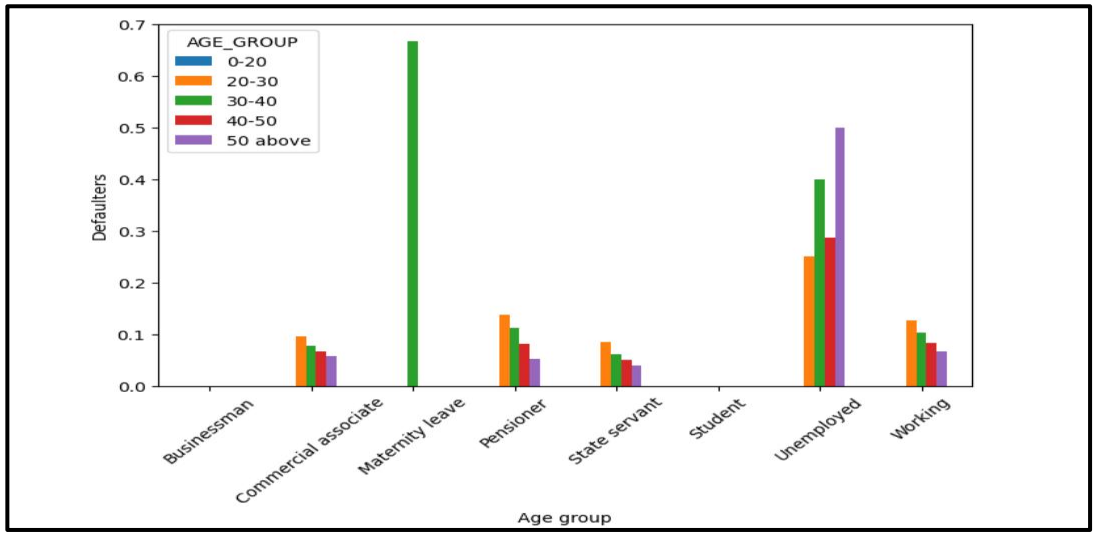
- **Data Analysis:-**
  - Income Type & Gender:



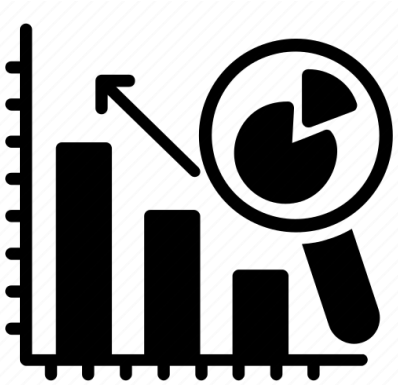
- Unemployed clients along with clients with females who are on maternity leave are heavily defaulted.
- It advisable to avoid giving loans to unemployed clients and clients who are on maternity leave.

INSIGHTS

- **Data Analysis:-**
  - Income Type & Age Group:

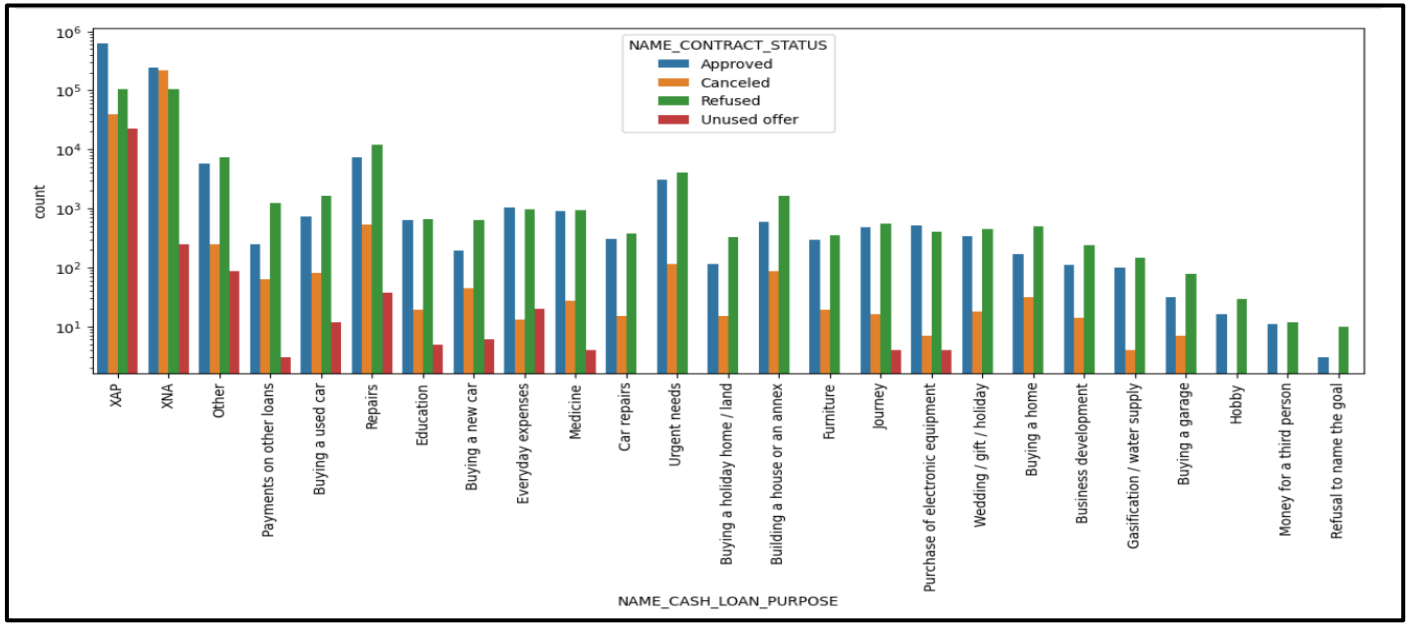


- Middle age groups and senior client are less likely to defaulted.

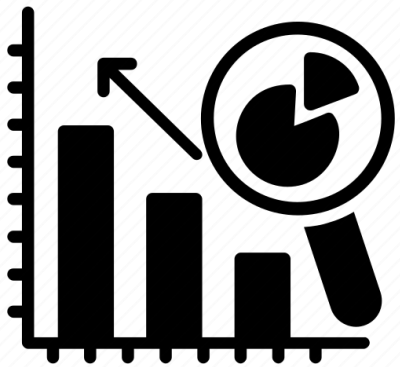


INSIGHTS

- **Data Analysis:-**
  - Analysis on Merged dataset:

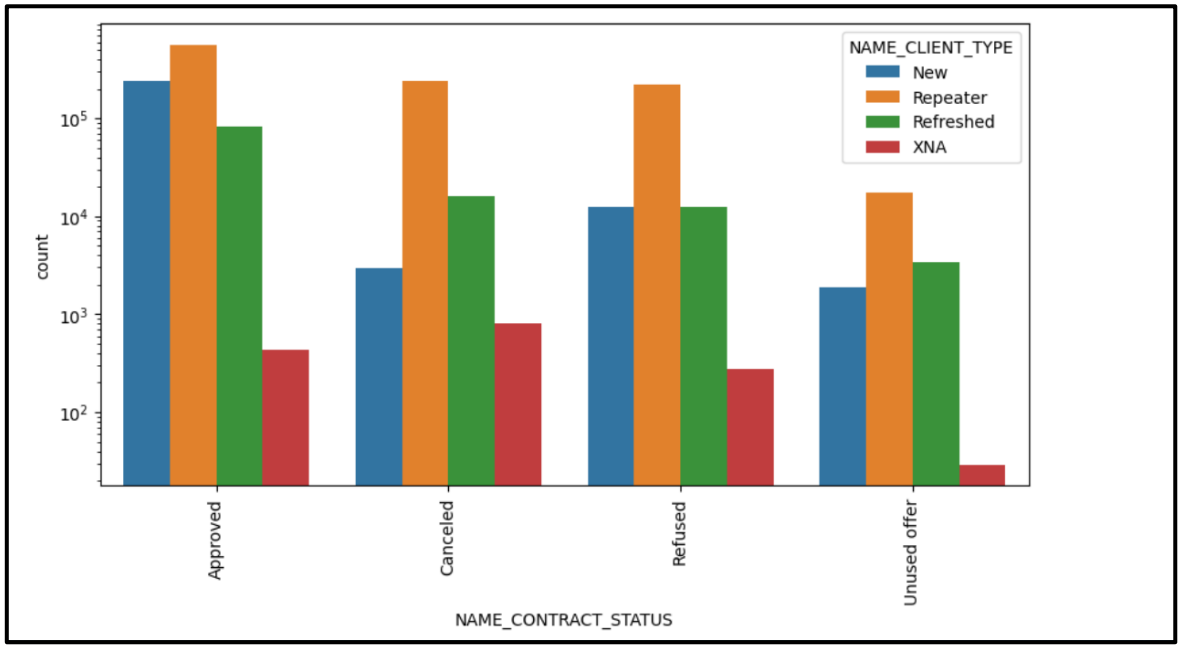


- For the repairing purpose customers had applied for loan previously and the same purpose has most number of cancellation.

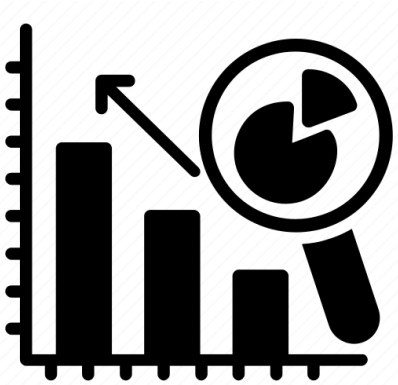


INSIGHTS

- **Data Analysis:-**
  - Analysis on Merged dataset:

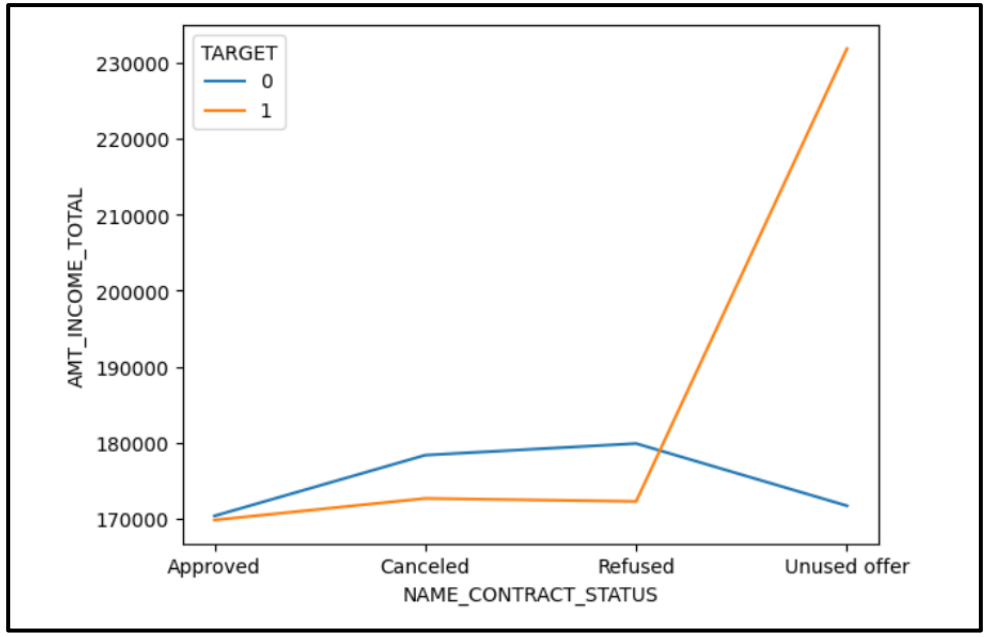


- There is a risk to grant loans for clients, whose applications were refused or unused previously.

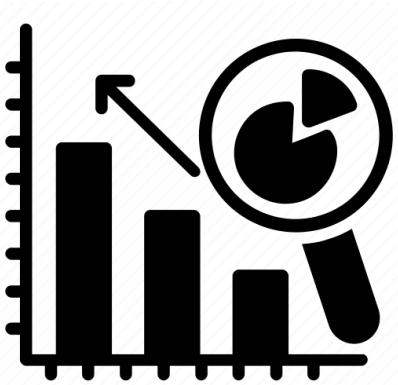


INSIGHTS

- **Data Analysis:-**
  - Analysis on Merged dataset:

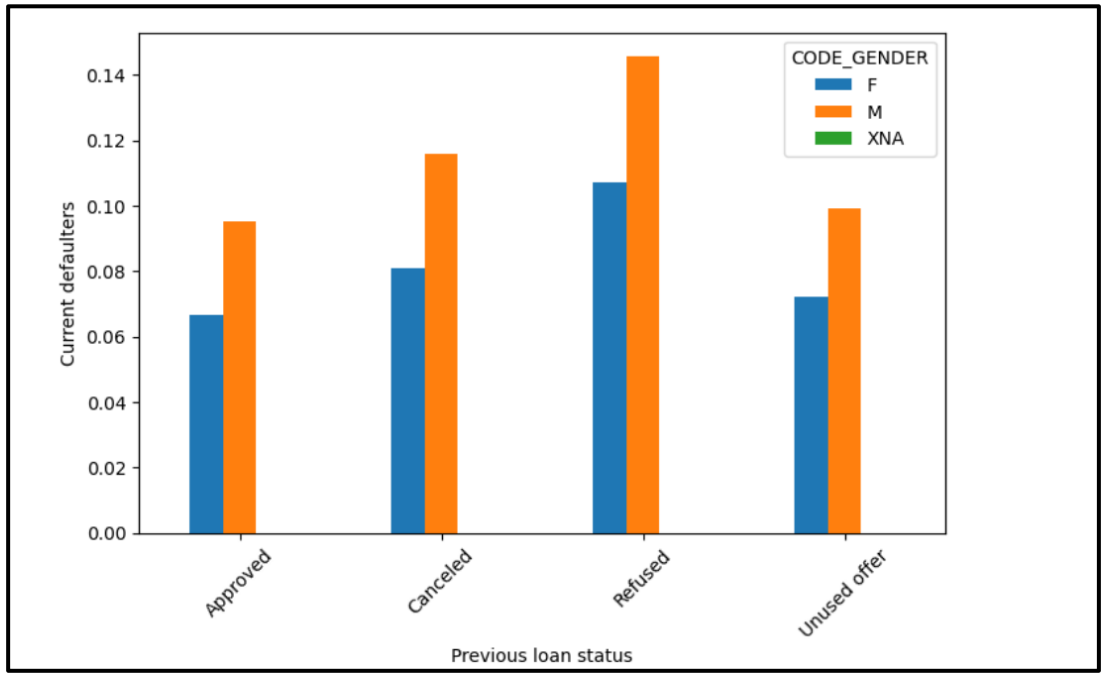


- Offers which were unused previously now have maximum number of defaulters despite of having high income band customers.

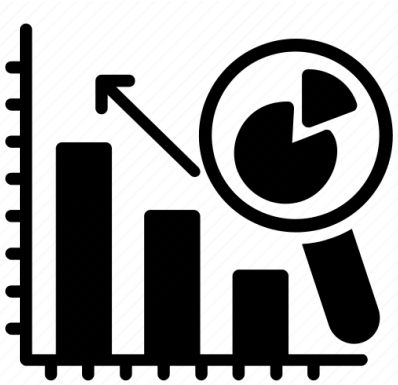


INSIGHTS

- **Data Analysis:-**
  - Analysis on Merged dataset:

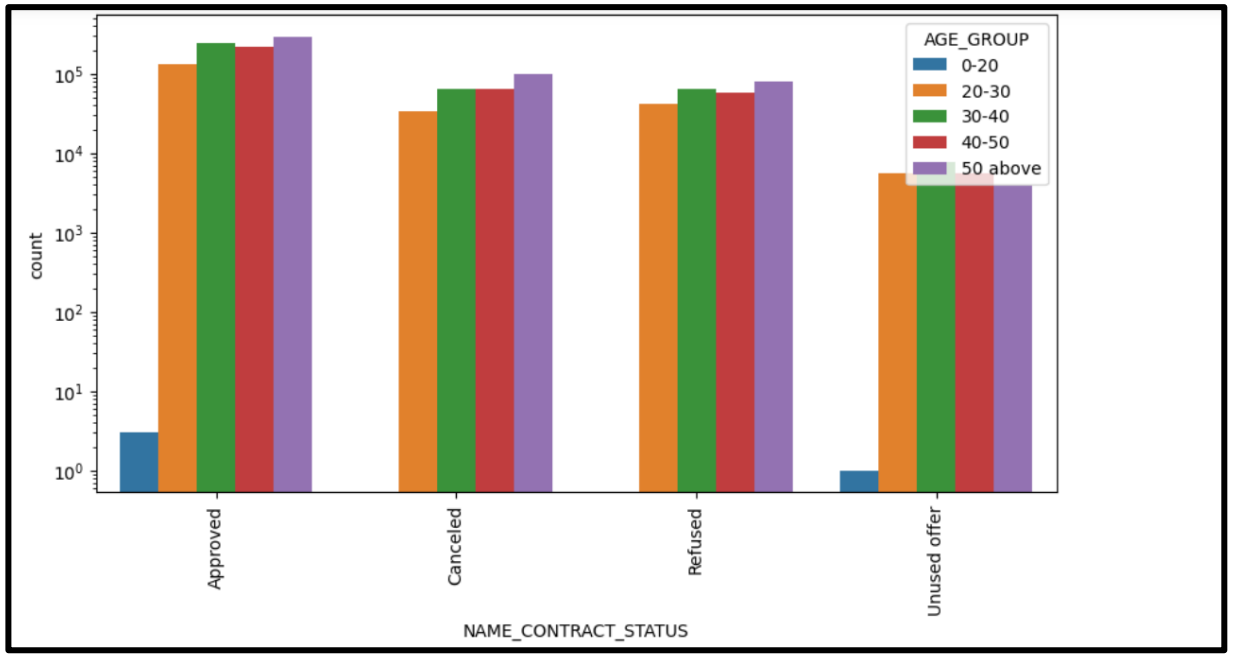


- Previously refused and unused offer applications were more defaulted in male.
- It is recommended to provide loans to previously approved females.

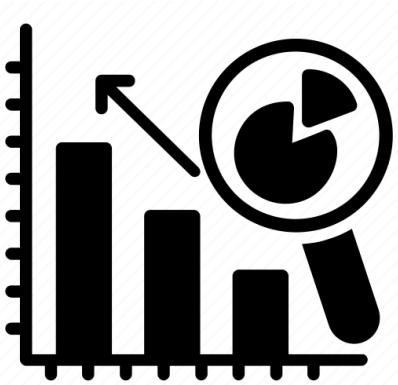


INSIGHTS

- **Data Analysis:-**
  - Analysis on Merged dataset:



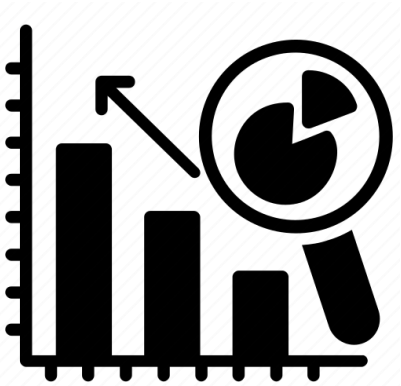
- Young people, who were previously refused are mostly defaulted.
- Safer to grant loans for senior citizen as they are less defaulted irrespective of their previous loan status.



INSIGHTS

CONCLUSION:

- It is recommended that bank should grant loan to highly educated clients with higher income.
- It is risky to give loans to clients who have low income with previously refused status.
- Unemployed customers are risky to give loan.
- Organization Transport Type 3 should be avoided.
- Low-Skill Laboreres and driver should be avoided.
- It is recommended to give preference to married clients.
- Young people are riskier segment to grant loan.



RESULTS

From these project I understood the importance of EDA and how can we derive meaningful insights or trends from data for a business solution. I got to learn Python and Python's libraries which are important for data analysis such as matplotlib, pandas, numpy and so on.

Drive link:
https://drive.google.com/file/d/1hWqQpFKPwSbliOK2cAU7kl0nA_YqQtY9/view?usp=sharing

THANK YOU