CREDIT EDA CASE STUDY

-BY VIJAYA SALIKAR AND SHRUTI PAWAR

Case Study Objective

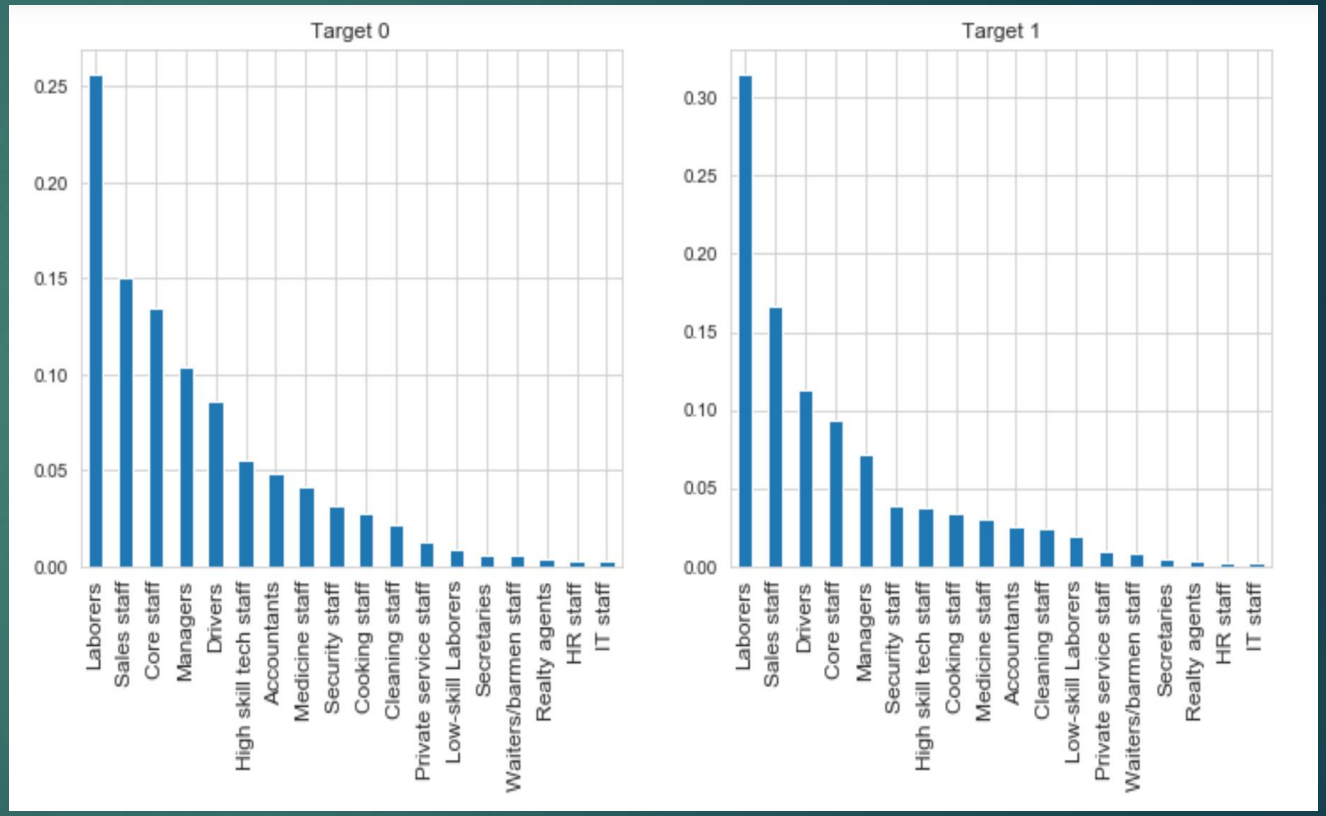
- To analyze the Application and Previous application dataset of clients applying for loans.
- Identify potential customers for loans.
- Help the bank make decisions if customer should be given loan so that bank can maximize the profits through interest earned and minimize on defaulting customers to avoid losses
- These variables would be identified using Exploratory Data Analysis Approach on the dataset.

Steps Followed to identify the variables in this case study

- Application_data.csv was read and inspected for missing values, outliers , data types and cleaned / analyzed for missing value treatments .
- Application_data identifies the customers who are defaulters and non-defaulters based on a column called as Target variable which has 2 distinct values -1 indicating customer is defaulted , 0 – indicating customer is a non-defaulter
- Further analysis was done with respect to this “Target” column and inferences were drawn.
- Previous application dataset was analyzed with respect to the loan approval status of the customer to determine other variables in decision making

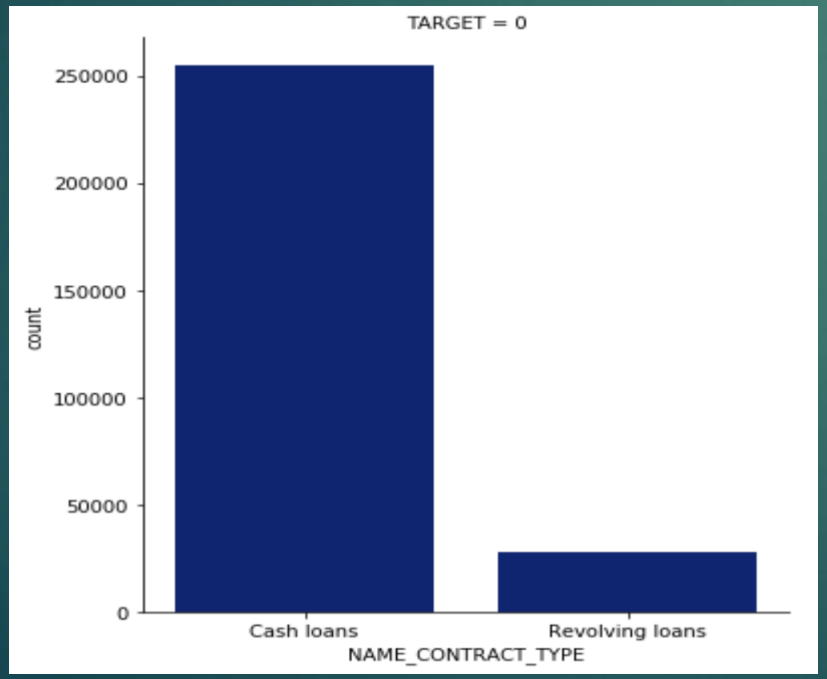
Univariate analysis for Target 0 and Target 1 categorical Variable - OCCUPATION_TYPE

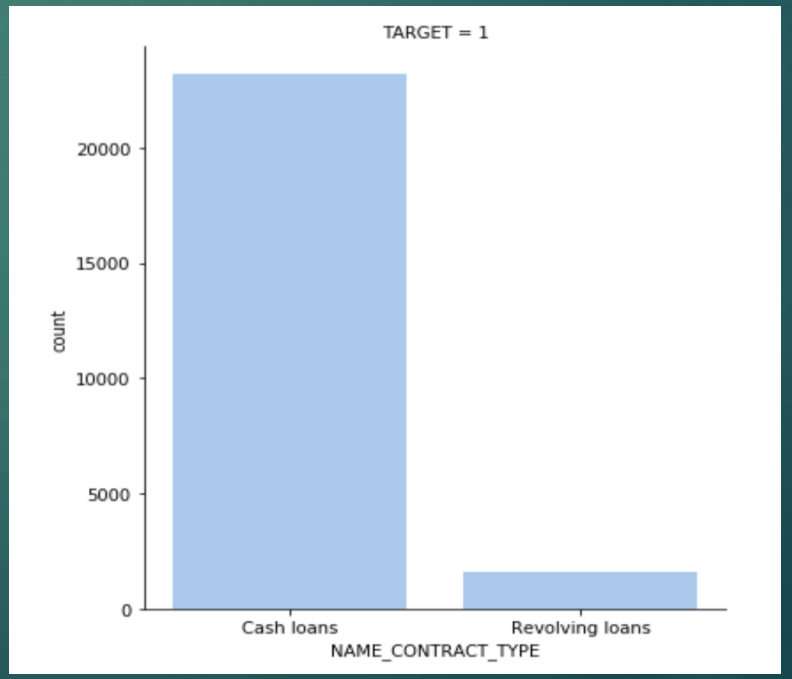
Customers who are defaulters(Target 1) and non-defaulters (Target 0), belong to occupation type "laborers“. They constitute a high percentage of applicants as compared to other occupations and the lowest is for IT staff.



Univariate analysis for Target 0 and Target 1 categorical Variable - NAME_CONTRACT_TYPE

As seen from the above graphs we see that, in both defaulters(Target 1) and non-defaulters (Target 0), customers have high number of applications are for Cash Loan as compared to Revolving Loans



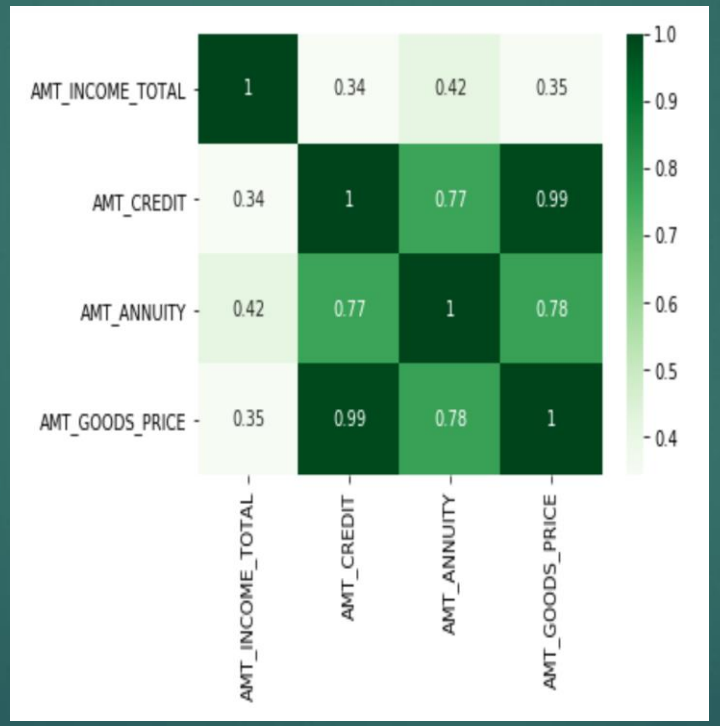


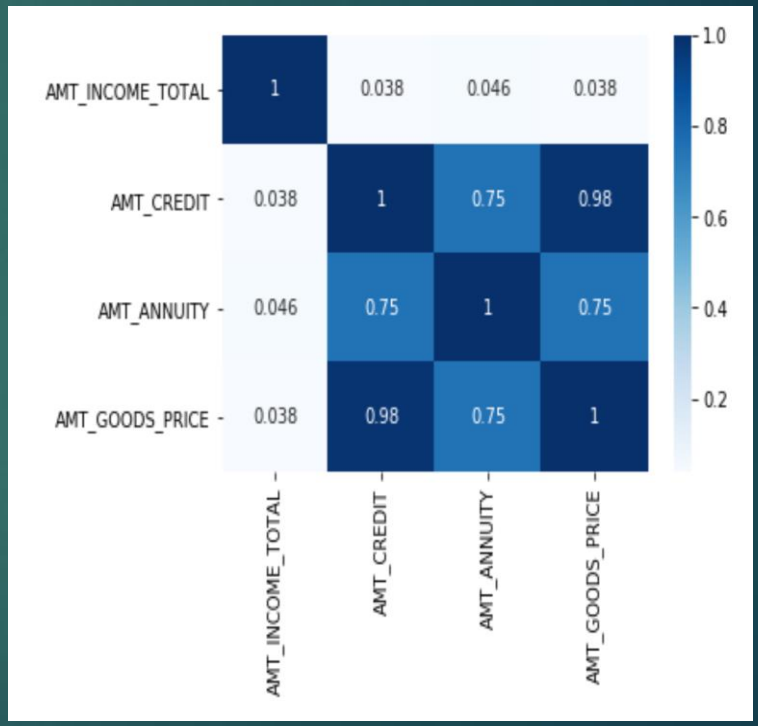
Correlation of numeric columns for both cases Target 0 and Target 1

1) From the above 2 heatmaps for target 0 and target 1 , we can see that variables AMT_CREDIT and AMT_GOODS_PRICE are highly corelated in both the data frames which is around 0.99

2) AMT_ANNUITY, AMT_GOODS_PRICE are medium correlation in both target 0 and Target 1 which is around 0.75

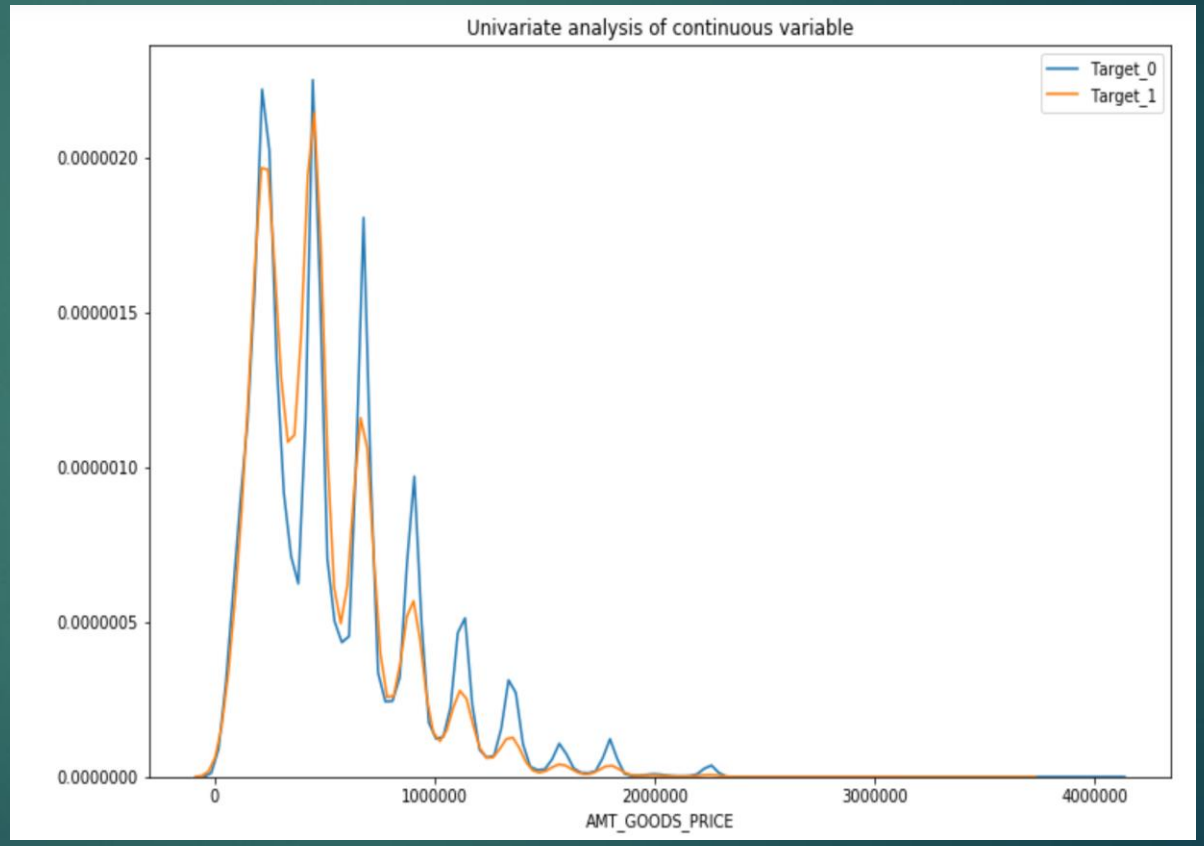
3) AMT_INCOME_TOTAL is the least corelated variable





Analysis for Target 0 and Target 1 for AMT_GOODS_PRICE

We observe that People who are defaulters (Target 0) seem to have higher price of goods for which the loan is given as compared to Customers who defaulted (Target 1)



Credit Amount and Occupation of client for TARGET 0 and TARGET 1 customers

For target 0

1) Both Male and Females belonging to occupation of "Managers" have a higher credit amount as compared to other occupations

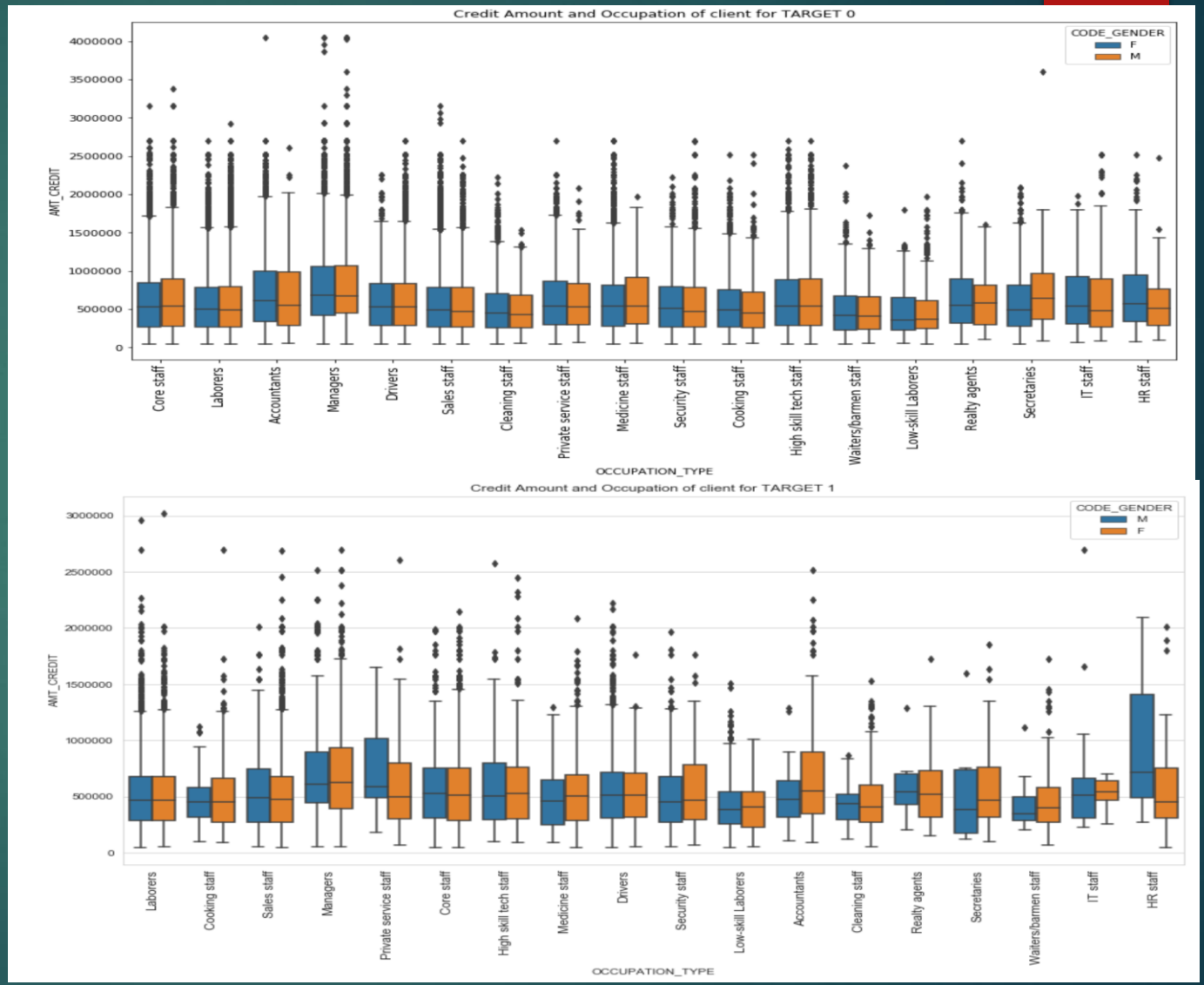
2) Both Male and Females belonging to occupation of "Low skill laborers" have the lowest credit amount as compared to other occupations

For target 1

1) Males belonging to occupation of "HR staff" have a higher credit amount as compared to other occupations

2) Females belonging to occupation of "Accountant" have a higher credit amount as compared to other occupations

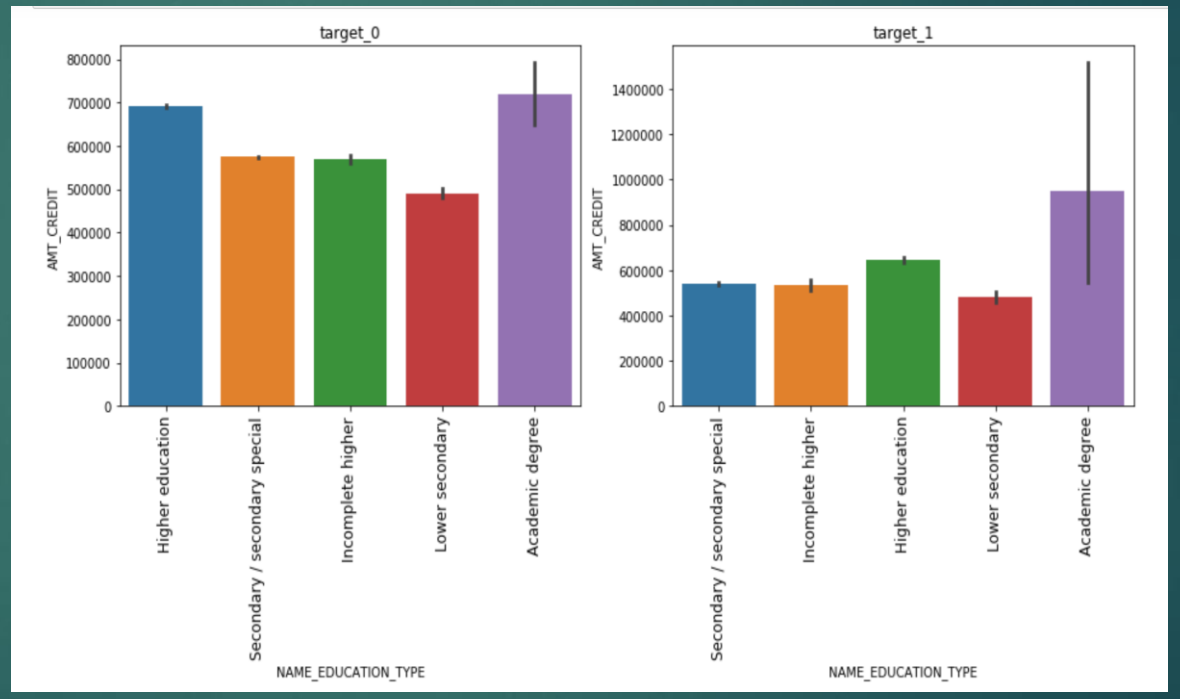
2) Both Male and Females belonging to occupation of "Low skill laborers" have the lowest credit amount as compared to other occupations



Education Type and Amount of Credit Relation

1) For Target 0 and Target 1,customers who have an academic degree or have a higher education have a higher credit amount of loan

2) For Target 0 and Target 1,customers who have a lower secondary degree have the lowest credit amount of loan



Various categories of Cash Loans applied as per Gender

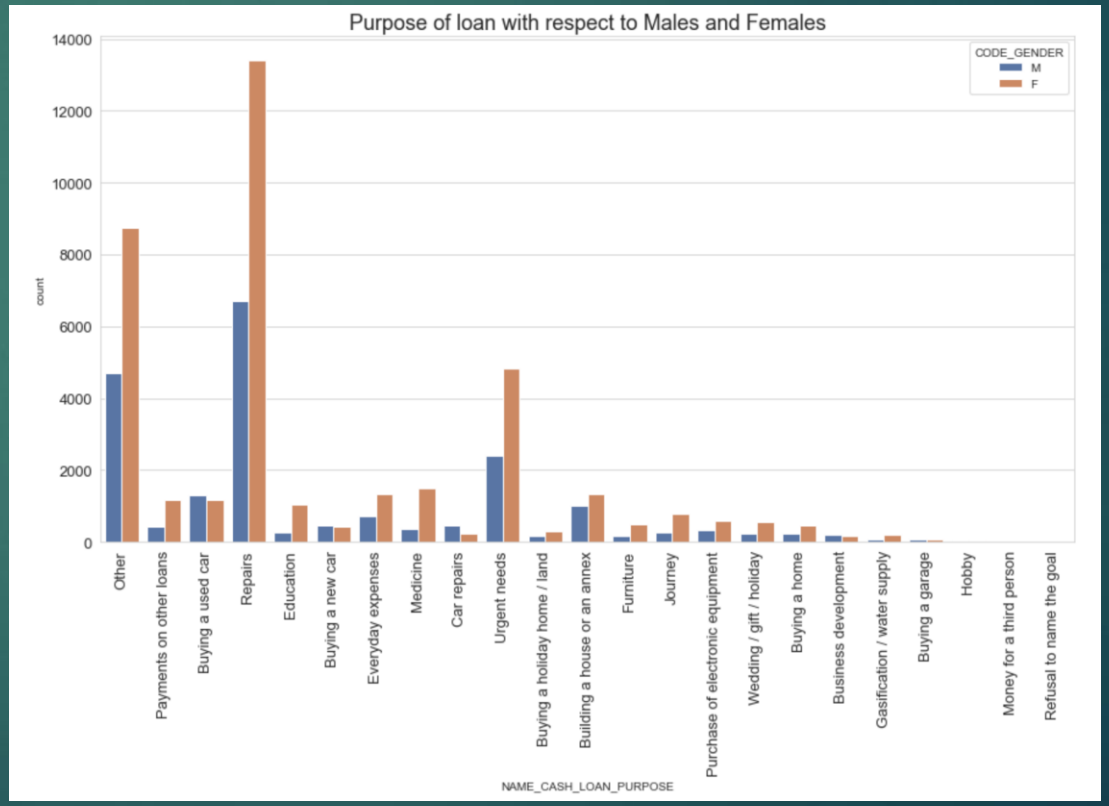
1) Highest number of applications are for Repair purposes and the count is applied by Females.

2) In the "Urgent Need " category the number of applications for females is double of Males

3) In the categories related to Cars - buying/Repairs/ the number of applications made by Males is higher than females.

4) In the "Medicine" category the number of Males applying is very less in comparison with Females

5) The least applied category for loans is " Buying a Garage"



Approvals based on Various channels of Client Acquisition

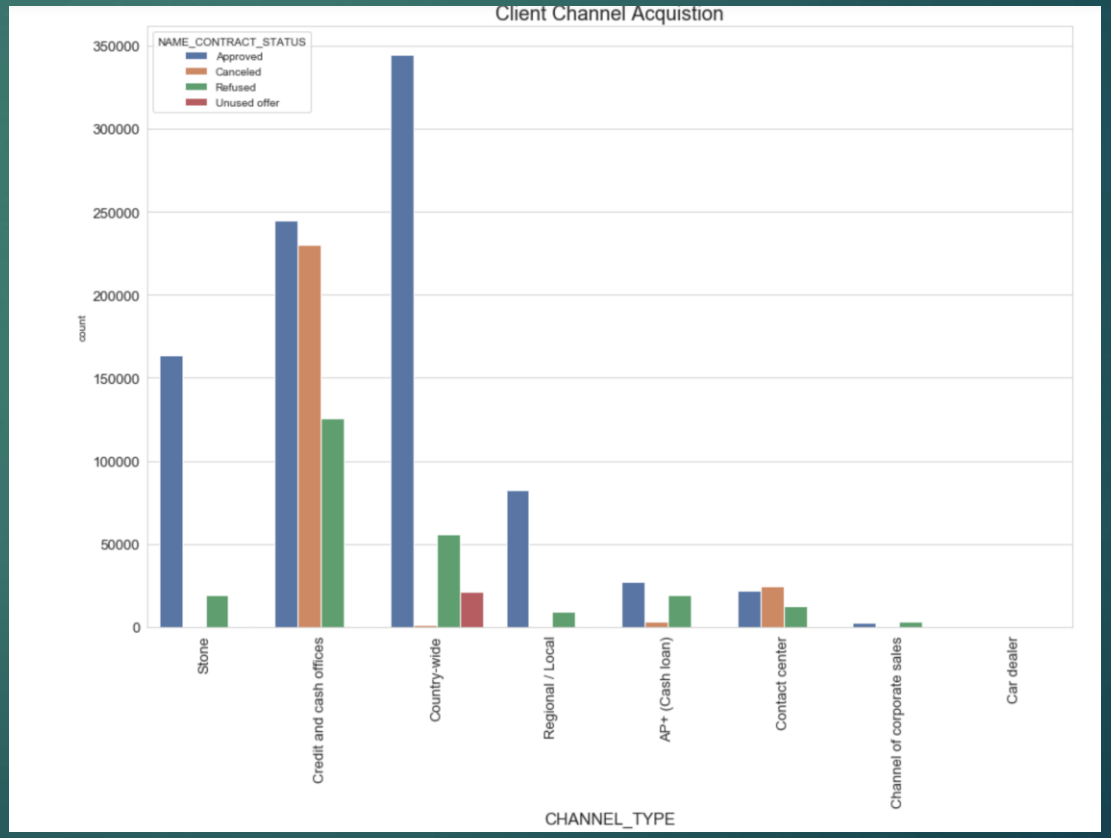
1) Highest number of acquisitions were through "Country Wide" channel and the number approval rate is also the highest for this and there are very few cancellations

2) Acquisitions made through "Credit and Cash offices" channel have the highest number of rejections

3) Country-Wide people have a high amount of unused offers -they can be identified as potential customers for loans

4) The "channel of Corporate Sales" does not seem to be highly tapped and can be focussed on to increase the client base

5) The clients approaching through " Credit and cash offices" see the highest refusal rate



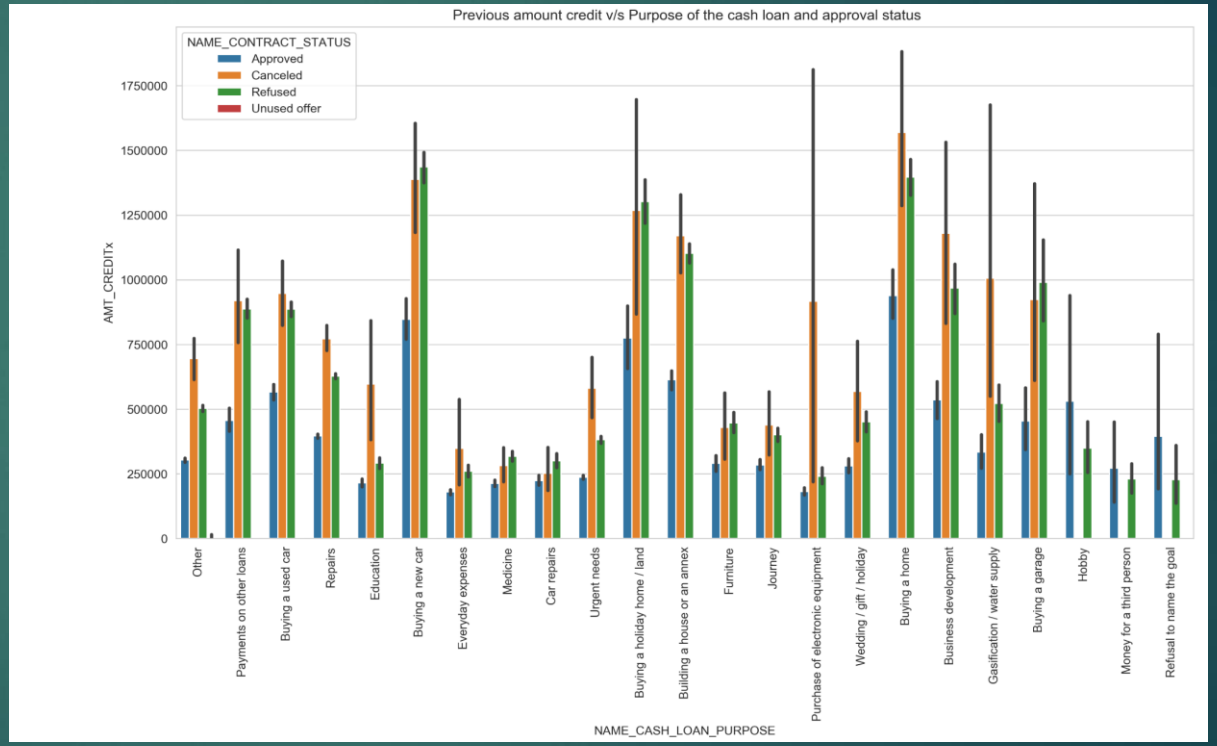
Approvals based on Previous Credit Amount and Loan Purpose

1) The highest amount of credit loan approved and Cancelled was for "Buying Home"

2) The highest amount of credit loan refused was for "Buying a new Car"

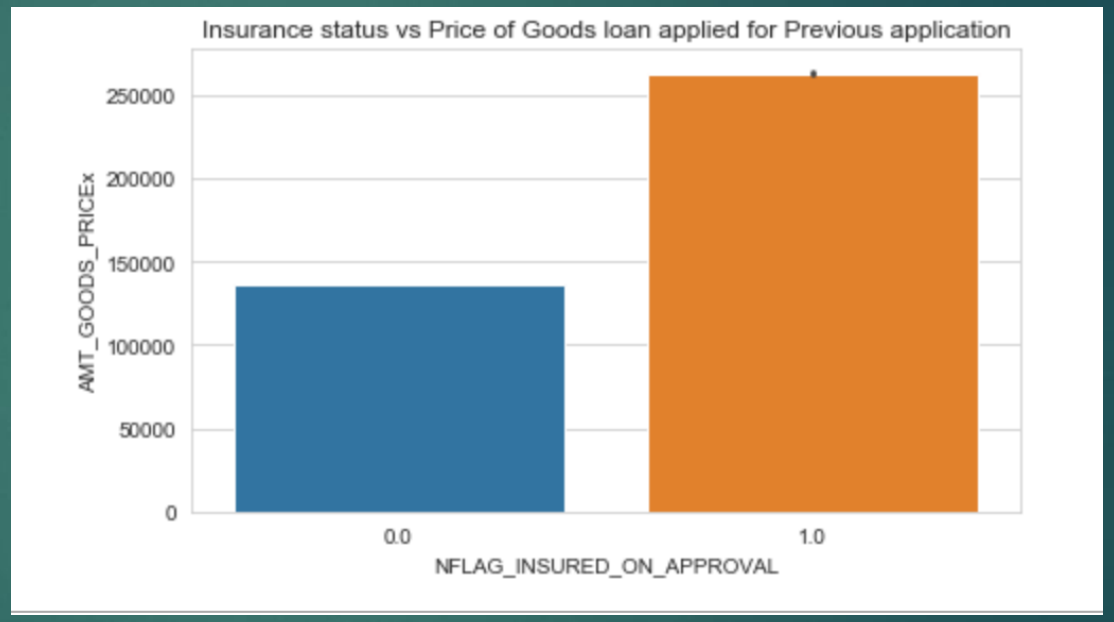
3) The lowest amount of credit loan approved is for "Purchase of electronic equipment"

4) The applications who have a credit amount in the range of 0 to 1000000 have got approved, applications beyond the 1000000 have got refused across all categories



Price of Goods for loan applied V/s Insurance Status

We see that if the amount of goods price is higher then , clients have taken insurance



Conclusion

1. The main purpose of the case study is to classify and analyze the nature of the loan applicants.

2. From a proper analysis of data set and constraints of the banking sector, various different graphs were generated and visualized.

3. From the graphs, many conclusions have been made and information were inferred.

4. The cash loan was preferred over revolving loan by majority of the loan applicants from both male and female category.

5. Customers majorly apply loans in the category of "Repair" followed by "Other purpose" and "Urgent needs only".

6. Low and moderately categorized applicants try for .

7. Clients who own their own apartments should be preffered over clients who stay in rented apartments so the bank can expect successful payments and minimise on the number of defaulters.

8. The bank should concentrate on customers having a higher annual income for loans from the non default category for timely payments.

9. The Client prefer taking the insurance from bank if the price of goods for which loan is applied is high.

10. The clients approaching through " Credit and cash offices" see the highest refusal rate.Country-Wide people have a high amount of unused offers -they can be identified as potential customers for loans.

11. This case study work can be extended to higher level in future , Predictive model for loans that uses machine learning algorithms, where the results from each graph of the case study can be taken as individual criteria for the machine learning algorithm.