**CASE STUDY : EXPLORATORY DATA ANALYSIS ON LOAN PROVIDING COMPANY**

**INTRODUCTION:**

This case study is a part of CDAC End Module Exam.

To reach the conclusion and answer key we will go through the following phases of data Analysis process :

* Ask
* Prepare
* Process
* Analyze

**SCENARIO :**

In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

*Business Understanding* :

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You 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.

When the company receives a loan application, the company has to decide for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision:

* If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
* If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

**Business Objectives**

This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.  The company can utilise this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

**STEPS Followed During the Analysis:**

1. **Ask-**

* *What is the problem trying to solve?*

**The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample,

**All other cases:** All other cases when the payment is paid on time.

* When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

1. **Approved:** The Company has approved loan Application
2. **Canceled:** The client canceled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
3. **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
4. **Unused offer:**  Loan has been canceled by the client but on different stages of the process.

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

1. **PREPARE :**

For the Analysis we have 2 dataset having previous\_application data, application dataset

* For the Case study we have used Python, Imported the dataset in pandas.
* To check the data integrity, we have to check the the consistency in column and data type.

1. **PROCESS :**

In this phase before any analysis we have to check the data consistencies by data cleaning process.

* Imported the dataset 🡪 application and previous dataset
* Drop the column which is having high numbers of null values, which are irrelevant to the Dataset.
* Checked the columns which are relevant to dataset.
* Checked the rows and columns of the dataset.
* Replaced the blank values with others for categorical data type
* Replace the blank values with mean for Numerical data type
* We have done some analysis on columns and there relationship with the target , keeping those columns which are more relevant .
* Checked if there is any negative values, days having negative values we have so we converted it to positive .

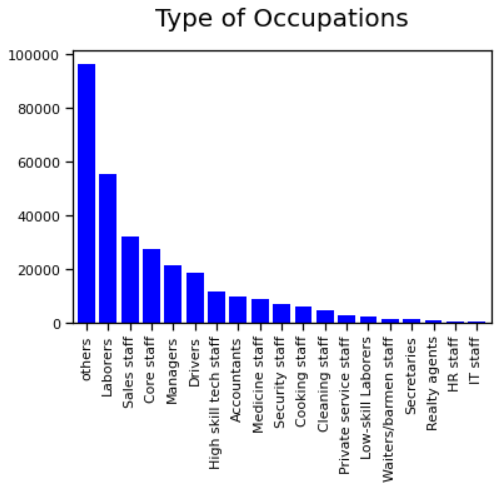
1. **ANALYZE :**

In this step of the Data Analysis Process, we will try to derive key insights from the cleaned data that we have gathered and use visualization techniques to get a better understanding of the story the data is trying to tell.

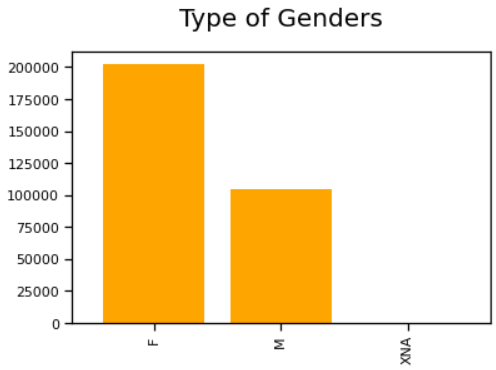
In this phase we have used pandas for visualization and finding the co-relation between the columns and how it affects the response.

Analysis For Application dataset :

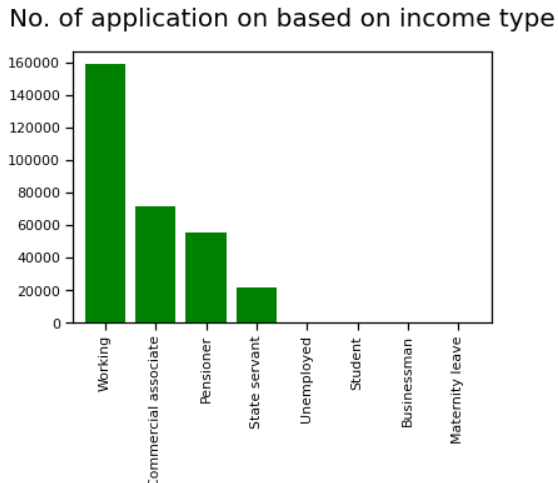
Graph 1: plotting graph on the basis of occupation



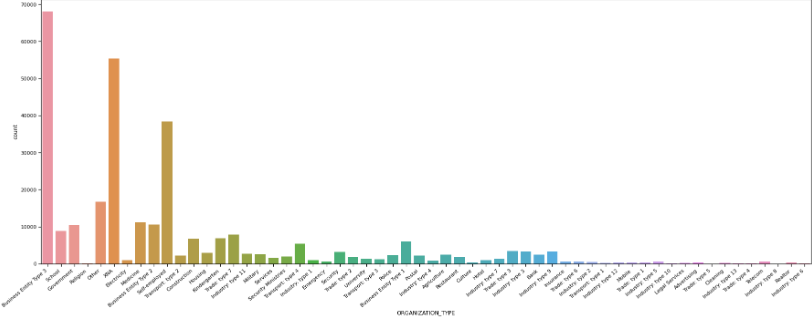
Graph 2:



Graph 3:



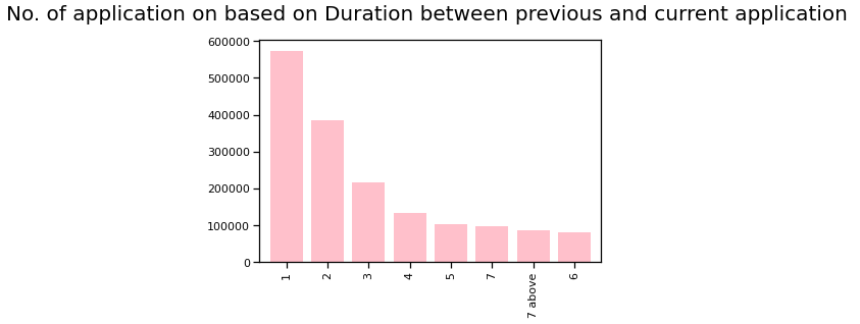
Graph 4 :



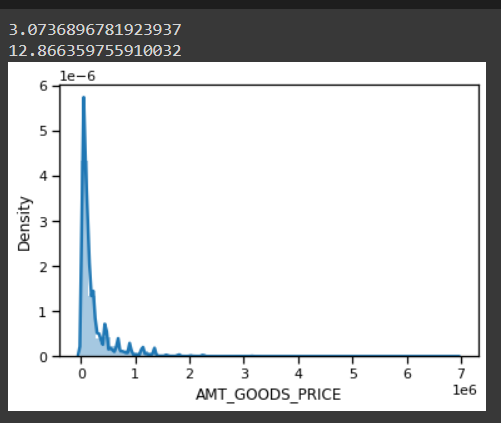
Identification of the outliers in Application dataset:

In this we have analyzed some of the columns having outliers.

Analysis on Previous Application :



Finding the kurtosis and skewness on the amt goods price :

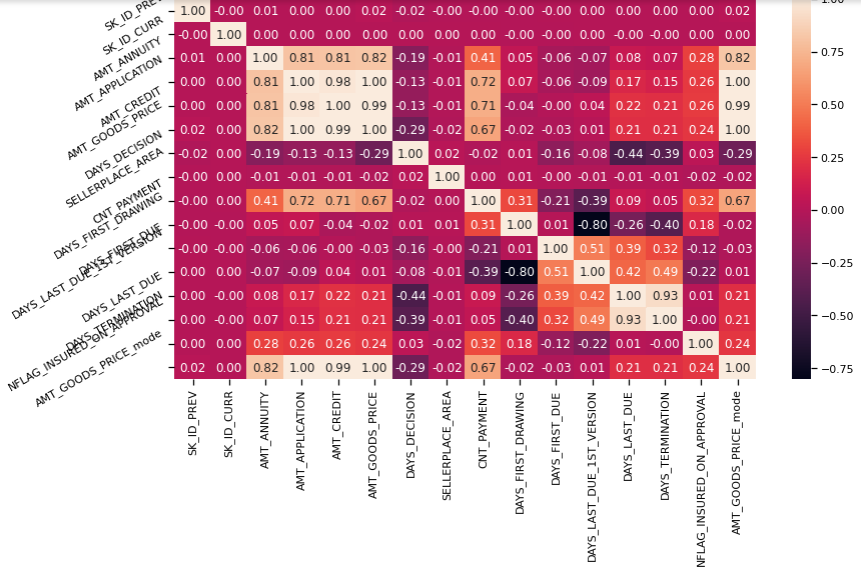


Identification of outliers :

In this we have analyzed some of the columns having outliers.

In this we have used box plot.

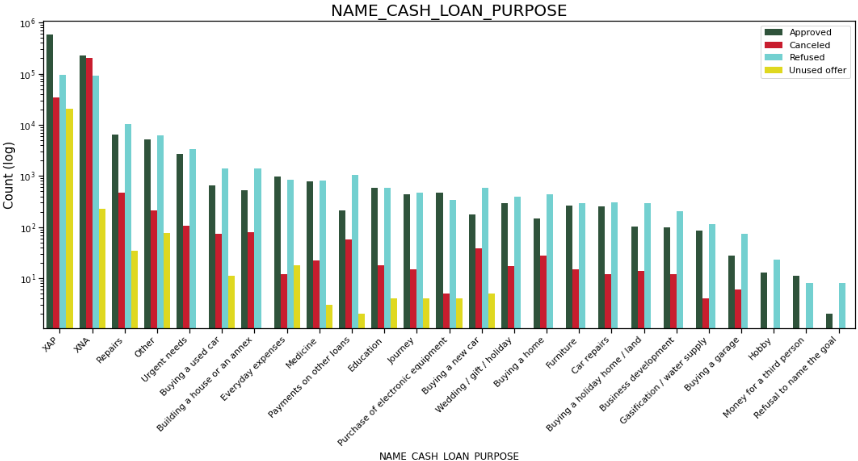
Finding the Co-relation on the coloumns🡪



**MERGING THE TWO DATASETS :**

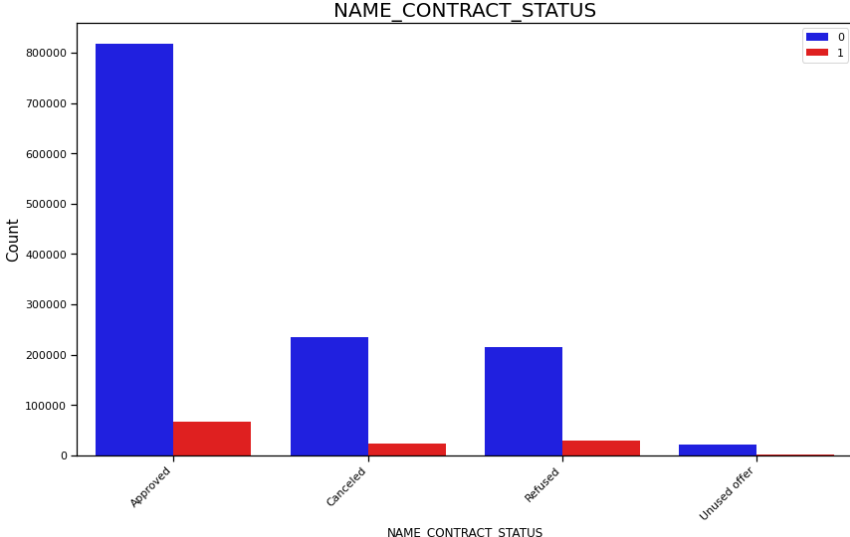
Graph 1 :

Checking univariant result for the merged data frame



Graph 2:

Analysis on the loan Contract status on the basis of previous loan repayment history .



Graph 3 :

Checking result for Inbalancing

