Credit EDA Assingment

Submitted By : Madhumita Roy

Google colab notebook link :
https://colab.research.google.com/drive/1d-kEeeC7RW3hwrmX6RNrB3ViVa9-kEq8?usp=sharing

Problem Statement

The loan-providing companies find it hard to give loans to 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 specialising in lending various types of loans to urban customers. You have to use EDA to analyse 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 on 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 for 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.

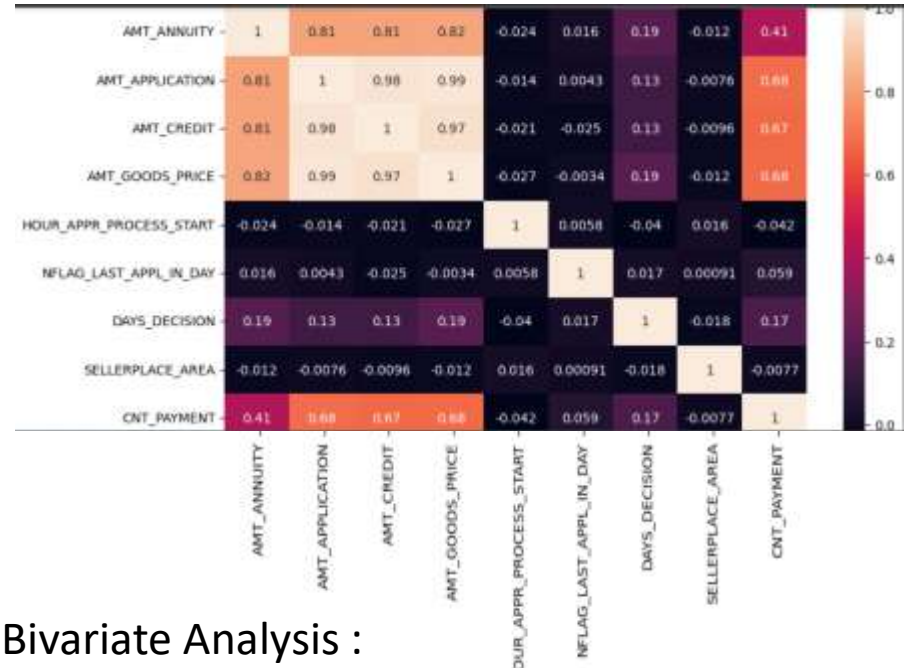
Approach and methodologies:

- To find the factors important for the sanction of a loan
- From the given three data sets 2 data sets are used for the analysis i.e. application_data and previous_application data as the third column column_sescription act as a data dictionary
- Then we import the library used for the analysis .
- After that we will load the data
- After that we will need to find out the columns with more than 40 % missing values and drop those columns because if in a data set some columns have more than 40% of its values missing than that means those columns are not really much essential for the analysis.
- Then again the data sets need to be checked for columns with less than 40% missing value , Now this columns can be necessary for the anaalysis so these colums need to be filled.
- To fill the missing values than the mean, mean, mode of all the columns should be found.
- After that the columns with missing values should be filled with their respective mean, mean, mode

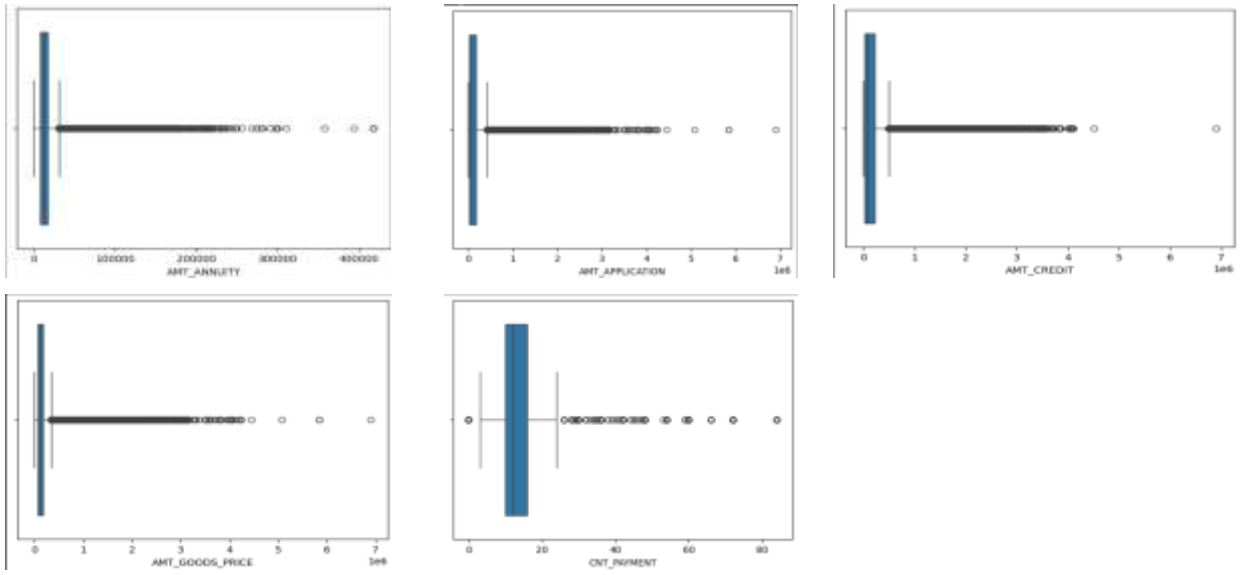
- The selection of the mean, mean, mode would be as such:
  for objects dtype : mode should be used to fill the columns
  for integer dtype : mean and median should be used depending on the necessity.
- For data whose mean and median are near in that case mean should be used as it makes the data more understandable and if the mean and median are far apart then the median should always be selected .
- After that the data should be divided into numerical and categorical data
- And on the basis of these data divisions graphs should be selected and plotted for univariet analysis .(For ex- box plot , hist plot, count plot etc.,)
- Then ouliers should be identified from the graphs and the reason why they are considered outliers should be stated
- Then the data imbalance for the target variable needs to be found.
- After that bivariate graphs should be on the basis of the necessity
- Then business insights for both the univariete and bivariete data needs to be found.
- After that heat map should be plotted for multivariate analysis
- Then top 10 most correlated colums should be found on the basis of the analysis
- After all this is over both the data sets needs to be merged together the same proocedure should be followed starting from the data cleaning to finding insights from univariate, bivariate and heat map .
- After that a ppt should be made stating the most important buisness insights along with the graphs.

Graphs and Insights

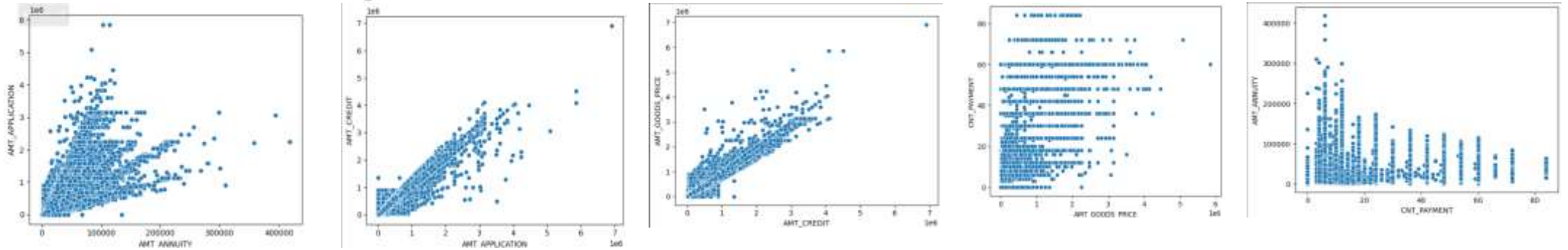
Heatmap :



Univariate Analysis :



Bivariate Analysis :



The most important factors which helps the lenders decide whom to provide loan are from the previous_application data are given below along with business insights for univariate analysis :

- AMT_ANNUITY :Loan products with annuities in the range 10000-20000 are likely well-received by clients, suggesting this is a competitive and acceptable range for the majority.
- AMT_APPLICATION : The demand for credit typically does not exceed 300,000 for the average population, indicating a preference or need for moderate loan amounts.
- AMT_CREDIT : Lenders have shown flexibility, often approving more than what clients initially asked for, indicating a high level of trust or favorable credit assessments.
- AMT_GOODS_PRICE : The credit that average of the population asked for in their previous application is around 120000
- CNT_PAYMENT : Clients prefer shorter loan terms, which can indicate a desire for quick repayment and less long-term financial commitment.

Business insights for bivariate analysis :

AMT_ANNUITY vs AMT_APPLICATION

Credit amounts and annuities are proportionally balanced, it suggests clients are taking loans within their repayment capacity, which can be seen as a lower risk.

AMT_APPLICATION vs AMT_CREDIT

When the final credit amount matches the requested amount, it suggests that the institution is meeting client expectations, which can lead to higher client satisfaction and loyalty.

AMT_CREDIT vs AMT_GOODS_PRICE

If the final credit amount (AMT_CREDIT) matches or exceeds the goods price (AMT_GOODS_PRICE), it indicates that the loan covers the cost of the goods entirely, suggesting a high approval rate for the requested amount.

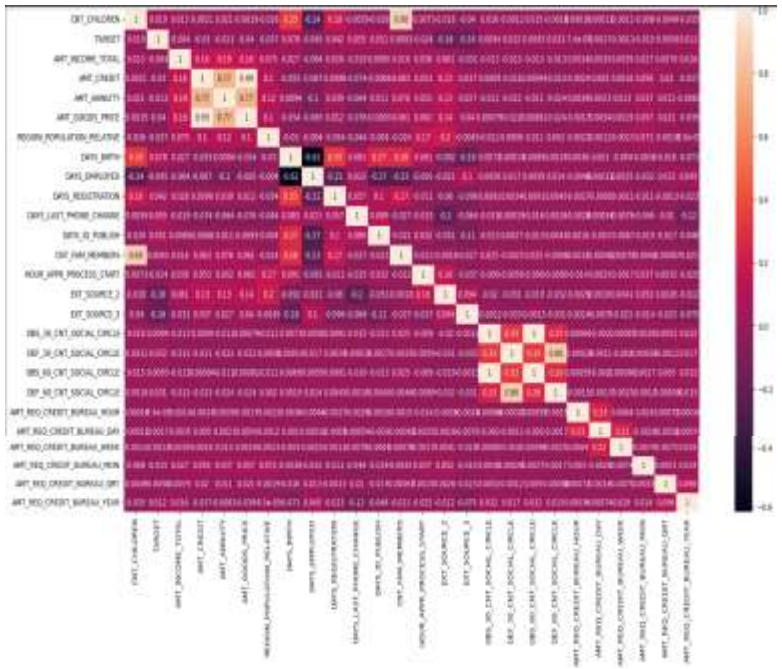
AMT_GOODS_PRICE vs CNT_PAYMENT

Clients requesting low-priced goods with short-term credit might indicate a preference for quick repayment and minimal long-term financial commitments.

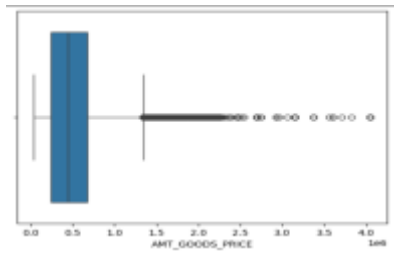
CNT_PAYMENT vs AMT_ANNUITY

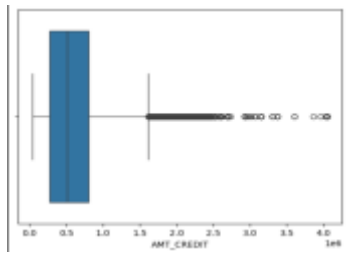
Longer terms with lower annuities can be designed for clients seeking lower monthly payments and vice versa.

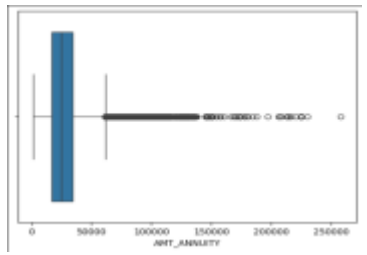
Heatmap :

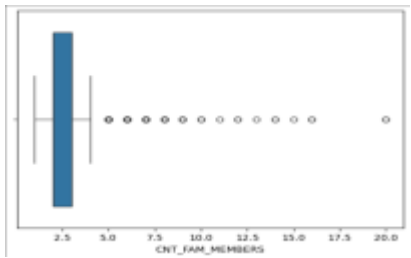


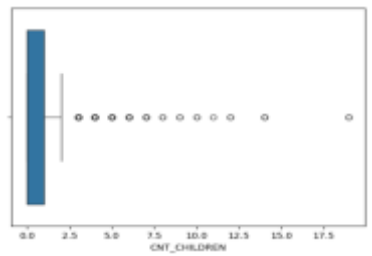
Univariate Analysis :



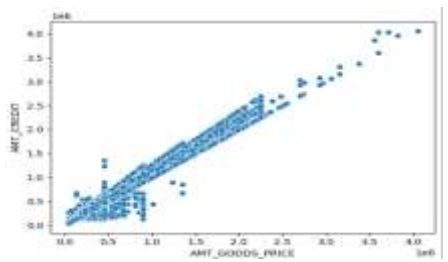


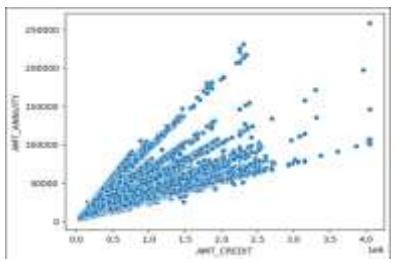


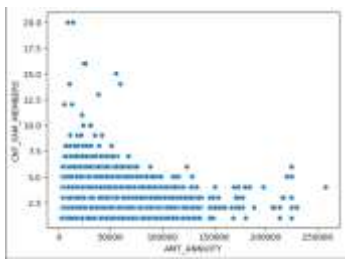


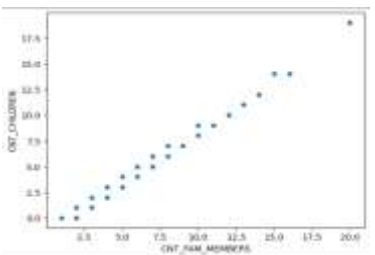


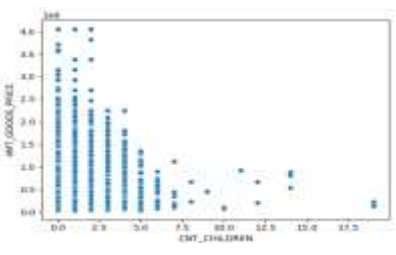
Bivariate Analysis :











The most important factors which helps the lenders decide whom to provide loan are from the previous_application data are given below along with business insights for univariate analysis :

- AMT_GOODS_PRICE : The range of goods prices falling between 250,000 and 750,000 indicates the typical affordability range for the majority of loan applicants. This insight is crucial for designing loan products that align with the purchasing power of the target market segment.
- AMT_CREDIT : The concentration of credit amounts between 250,000 and 800,000 suggests that this is the preferred loan size for the majority of borrowers. Lenders can use this insight to streamline their loan approval processes and focus their efforts on assessing loan applications within this range.
- AMT_ANNUITY : The range of loan annuities falling between 17,000 and 35,000 represents the typical affordability range for most borrowers. Lenders can use this insight to assess the borrower's ability to meet the monthly repayment obligations based on their income level.
- CNT_FAM_MEMBERS : The prevalence of households with 2-3 family members suggests that this is the typical household size among loan applicants. Lenders can use this information to tailor their loan products and services to meet the needs of borrowers with smaller or larger families.
- CNT_CHILDREN : The majority of clients have 0-2 children, indicating the typical family size among loan applicants. Lenders can use this insight to develop marketing strategies and loan products that cater to the needs of borrowers with children, such as education loans or family-friendly repayment options.

Business insights for bivariate analysis :

AMT_GOODS_PRICE vs AMT_CREDIT

By examining how closely the credit amounts align with the price of goods, lenders can understand client spending behavior and the types of goods being financed.

AMT_CREDIT vs AMT_ANNUITY

As the price of the goods increases, so does the loan annuity. This could indicate that borrowers are taking larger loans for more expensive purchases.

AMT_ANNUITY vs CNT_FAM_MEMBERS

This correlation between loan annuity and the number of family members suggests that borrowers with larger families tend to take out smaller loans. This could indicate that these borrowers have lower household incomes or prioritize budgeting more conservatively due to additional financial responsibilities.

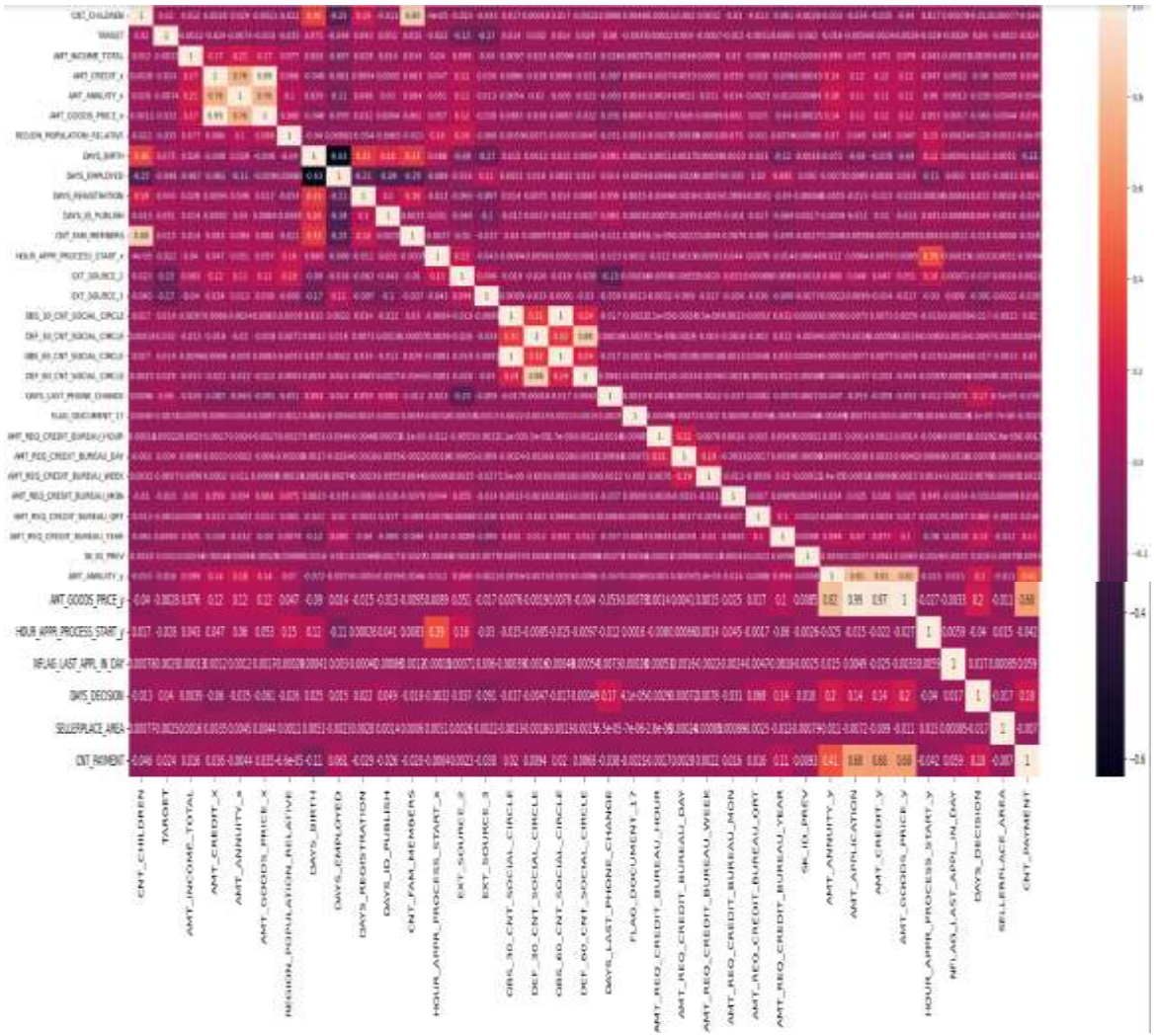
CNT_FAM_MEMBERS vs CNT_CHILDREN

The correlation between the number of family members and the number of children suggests that borrowers with larger families tend to have more financial responsibilities. This insight can inform lenders about the borrower's potential financial commitments and help assess their ability to manage additional debt obligations.

CNT_CHILDREN vs AMT_GOODS_PRICE

The correlation between the number of children and the price of goods for which the loan is taken suggests that borrowers with more children tend to opt for lower-priced goods. This insight can indicate that borrowers with larger families prioritize affordability and may be more budget-conscious in their purchasing decisions.

Heatmap :



Merged data heat map analysis
Gives the top 10 correlation same as that of the previous 2 datasets

Recommendation and conclusion

- Design loan products with annuities in the range of 10,000 to 20,000, as this range appears to be well-received by clients.
- Align credit amounts with the price of goods to understand client spending behavior and assess credit risk effectively.
- Meet client expectations by approving credit amounts that match or exceed the requested amount, leading to higher client satisfaction and loyalty.
- Develop marketing strategies and loan products that cater to the typical affordability range of borrowers, considering factors such as family size and the number of children.
- Offer financial planning support and resources to borrowers with larger families or specific financial needs, such as education loans or family-friendly repayment options.