CREDIT-RISK ANALYSER PROJECT

PRESENTED BY – ROCKY RANJAN

INTRODUCTION

- The two-dataset given were, ‘application_data’ and ‘previous_application’
- The problem statement was to conduct a comprehensive analysis of a dataset containing various financial and demographic attributes of loan applicants.
- Goal is to gain insights into the factors influencing loan default rates and to develop strategies to mitigate risks associated with lending.

DATA ANALYSIS PROCESS

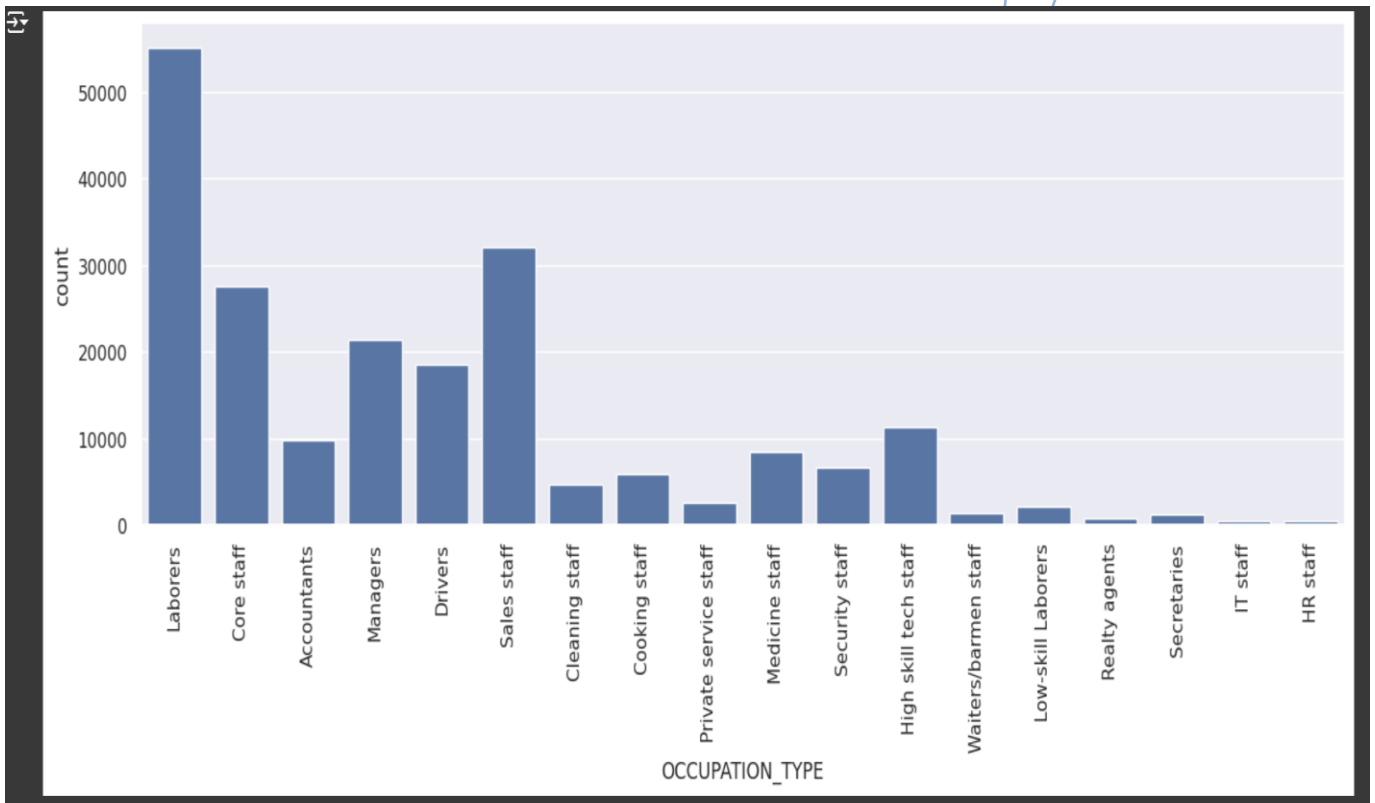
- Imported the necessary modules.
- As we can see there were multiple column with more than 60% of data missing, we dropped all of them.
- All the columns with less than 50% of and above the 32% of data were handled with either mean or median, depending on the distribution.

COMMONAREA_MEDI	69.872297
COMMONAREA_AVG	69.872297
COMMONAREA_MODE	69.872297
NONLIVINGAPARTMENTS_MODE	69.432963
NONLIVINGAPARTMENTS_AVG	69.432963
...	...
NAME_HOUSING_TYPE	0.000000
NAME_FAMILY_STATUS	0.000000
NAME_EDUCATION_TYPE	0.000000
NAME_INCOME_TYPE	0.000000
SK_ID_CURR	0.000000

122 rows × 1 columns

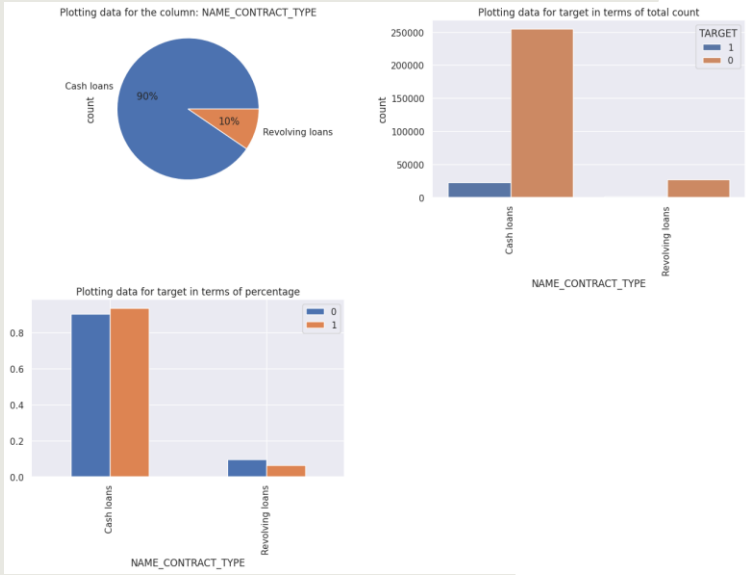
DATA ANALYSIS PROCESS

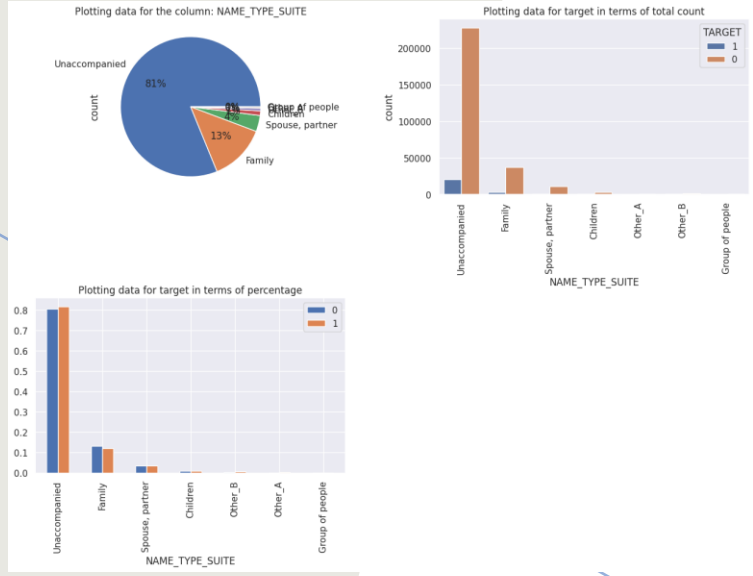
- The column “OCCUPATION_TYPE” was interesting, though it was an important column, the data missing were very large approx. 32%. I decided to fill in the missing data with “Unknown”



This column exhibits a significant proportion of missing values, amounting to 31%. Given that it is a categorical column, the only viable substitution for missing data is with the mode value. However, substituting all these missing values with the "Laborers" category wouldn't be appropriate. Therefore, we've decided not to perform any missing value treatment on this column and leave it as is.

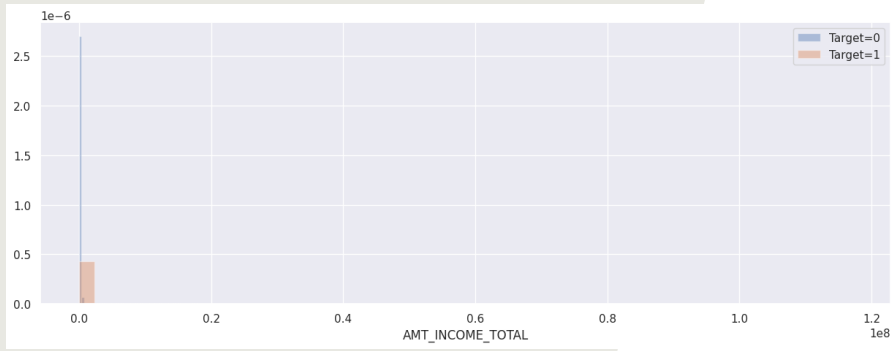
Note: Filling the missing values in the "OCCUPATION_TYPE" column by a new category named "Unknown"

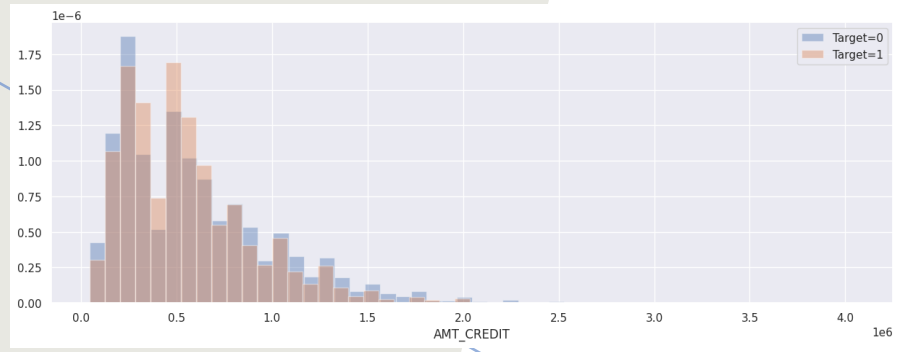




UNIVARIATE ANALYSIS(CATEGORICAL)

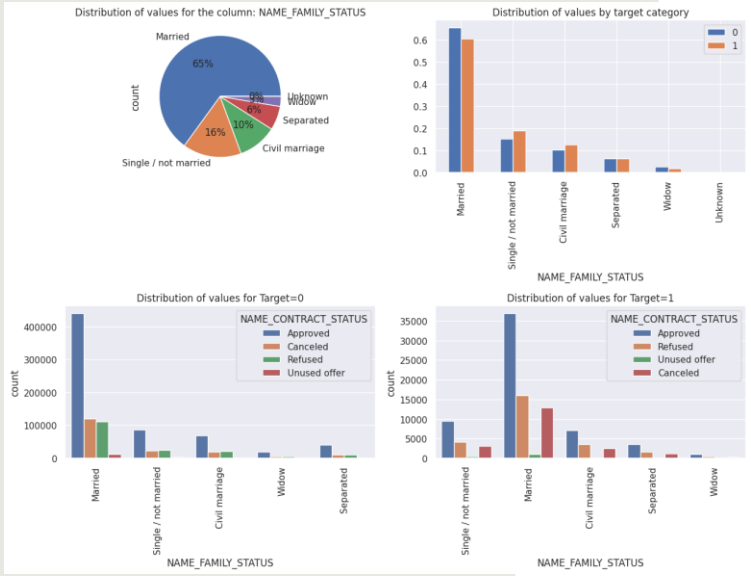
- Insights for NAME_CONTRACT_TYPE: The type of loan contract plays a critical role in predicting payment behavior. Cash loans may be associated with higher risk, possibly due to larger loan amounts or less flexible repayment terms compared to revolving loans.
- Insights for NAME_TYPE_SUITE : Applicants accompanied by family or a spouse during the loan application process appear to manage their repayments better, suggesting that household support might contribute to better financial behavior.

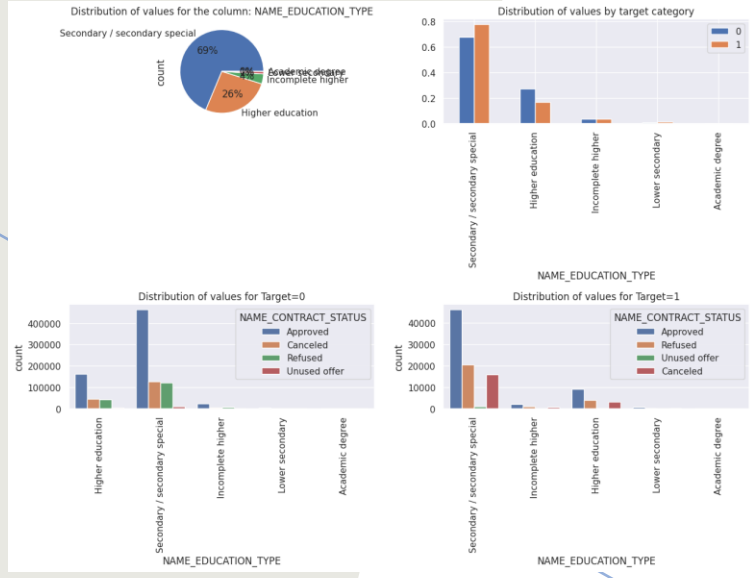




UNIVARIATE ANALYSIS(NUMERICAL)

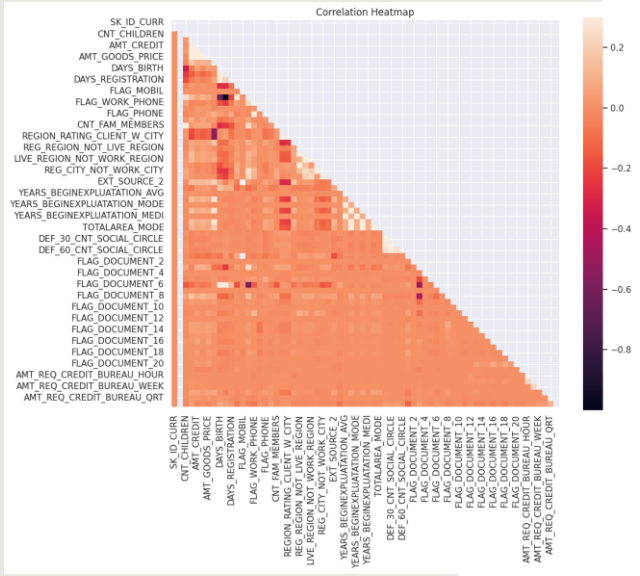
- Loan Amount Distribution: Smaller Loans (Up to 0.5 million): Individuals with repayment difficulties (Target 1) are more common in this loan size range. It indicates that people taking smaller loans may face more financial stress or difficulty in repayments.
- Income Levels: Lower Income and Repayment Difficulties: People in Target 1 (with repayment difficulties) have significantly lower incomes. Higher repayment difficulties are concentrated in the lower income brackets, making income level a strong determinant of repayment issues.

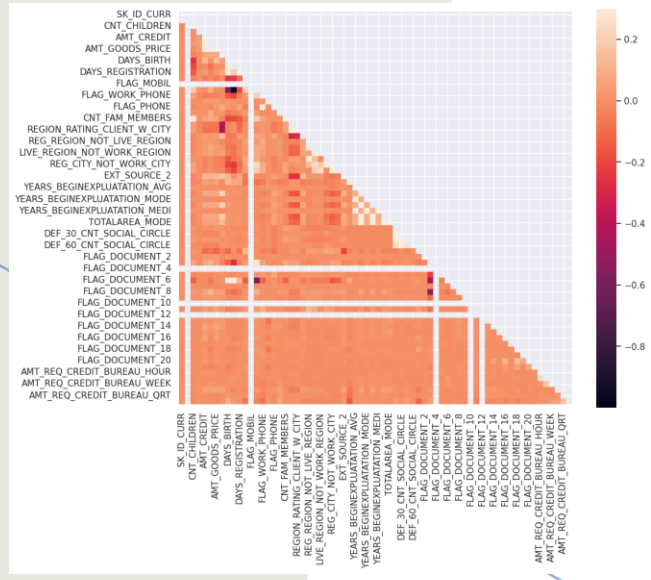




BIVARIATE ANALYSIS

- NAME_EDUCATION_TYPE : Education level plays an important role in the likelihood of experiencing repayment difficulties. People with Secondary/Secondary Special education are disproportionately represented among those with repayment difficulties (TARGET 1), and they are more likely to face canceled or refused loans.
- NAME_FAMILY_STATUS : These insights suggest that marital status plays a significant role in determining financial stability and repayment capacity. Single and Civil Marriage individuals are more likely to face difficulties, which could influence credit assessment criteria or inform the need for additional financial support programs targeted at these groups.



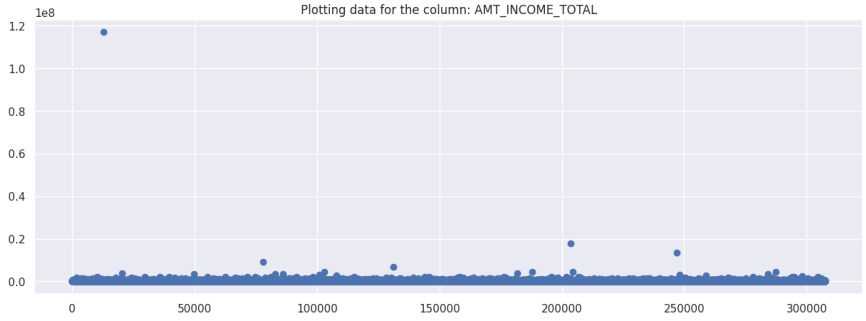


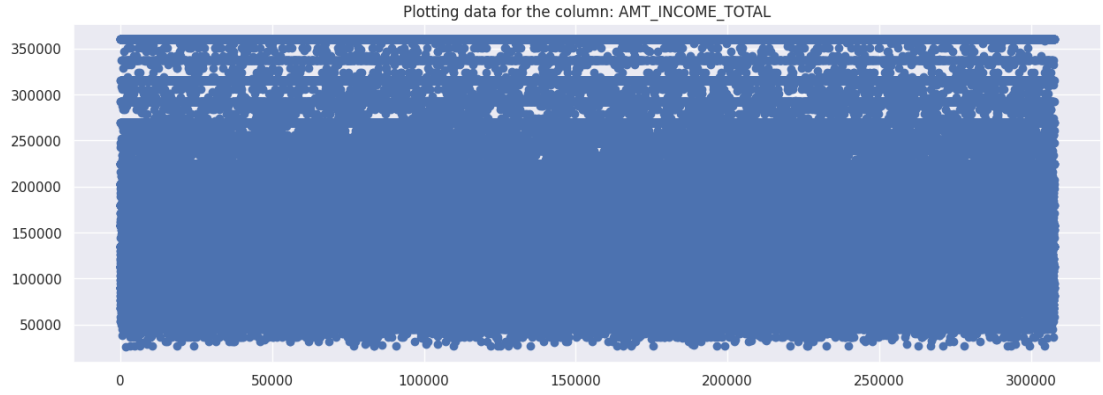
CORRELATION MATRICES

- These represent the most notable relationships found in the dataset. The relations with high correlation are primarily focused on financial indicators and population factors.

OUTLIERS

- Data Quality: Many of the outliers in the dataset seem to be data entry issues (e.g., DAYS_EMPLOYED, AMT_INCOME_TOTAL). These need to be corrected or excluded to avoid skewing any analyses or models.
- Extreme Cases: Some outliers, such as those in the OBS_30_CNT_SOCIAL_CIRCLE or AMT_CREDIT columns, could represent special cases (e.g., ultra-wealthy individuals or unusual social circumstances) and should be handled separately.
- Here is the image of AMT_INCOME_TOTAL with outliers





HANDLING OUTLIERS

- This is what I did for the outliers :
- Interquartile Range (IQR): This method identifies outliers as values outside 1.5 times the IQR below the first quartile (Q1) or above the third quartile (Q3).
- Capping: Instead of removing the outliers, we cap them at the lower or upper bounds, which can be more useful in maintaining data integrity.
- Here is the result of AMT_INCOME_TOTAL after handling outliers.

CONCLUSION

- A combination of income stability, credit scores, family size, loan size, and sector of employment provides a comprehensive framework for predicting repayment behavior. Lenders can use these factors to design more effective risk mitigation strategies and target borrowers who may require additional financial guidance or support.