Exploratory Data Analysis on Loan Application Data

IDENTIFYING PATTERNS IN LOAN DEFAULTS

OM PAWAR

Problem Statement

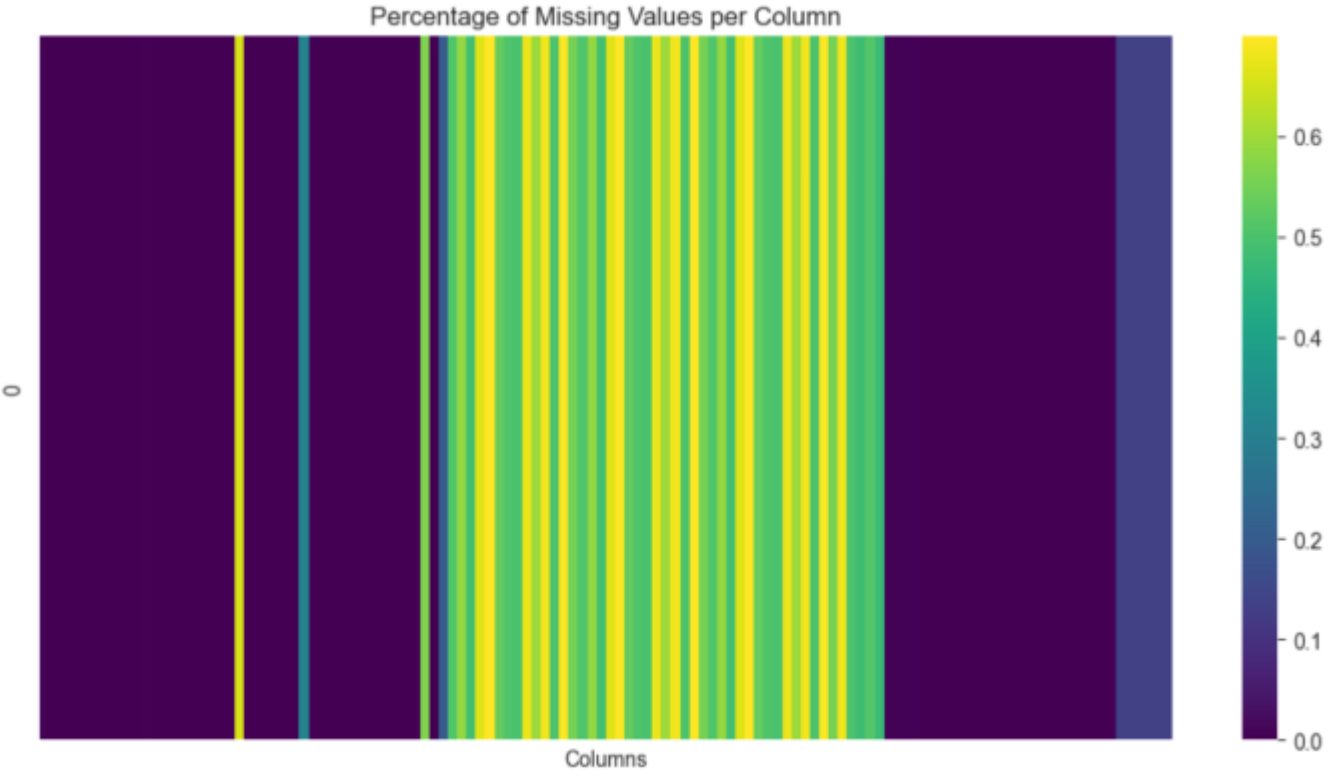
Conduct a comprehensive analysis of a dataset containing various financial and demographic attributes of loan applicants. Our goal is to gain insights into the factors influencing loan default rates and to develop strategies to mitigate risks associated with lending.

Data Overview

- application_data.csv file has 307511 rows.
- application_data.csv file has 122 columns.

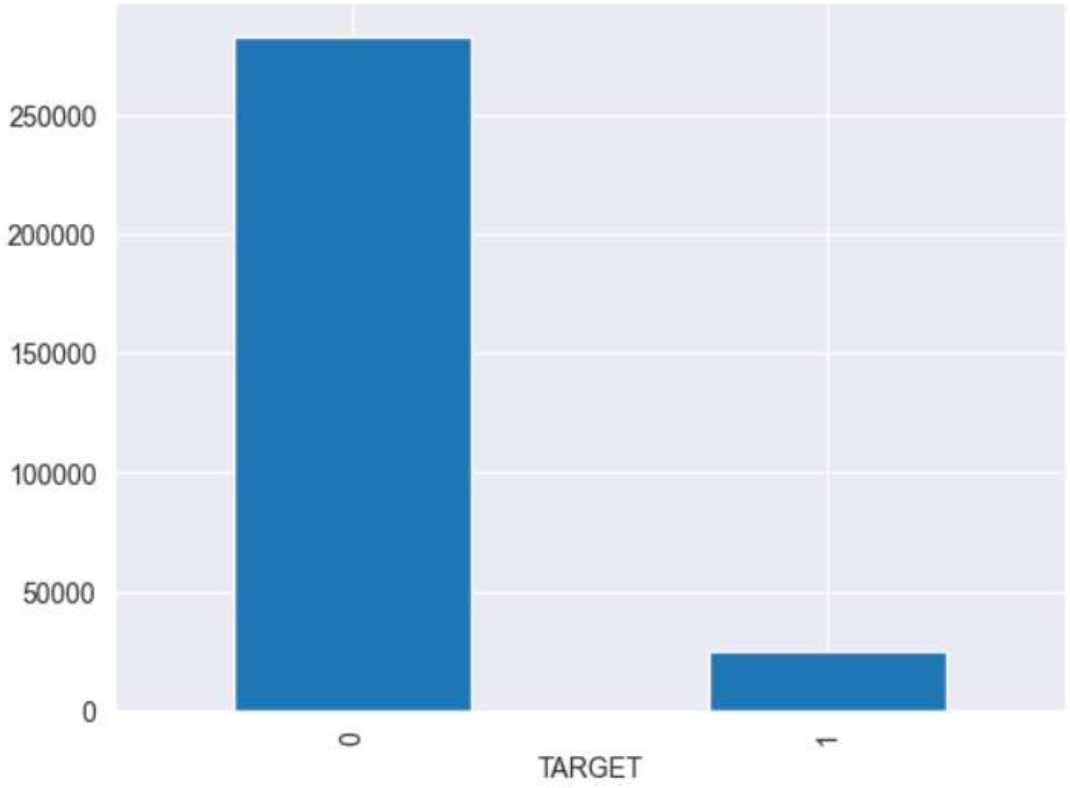
- previous_application.csv fila has 1670214 rows.
- previous_application.csv file has 37 columns

Missing Value Analysis



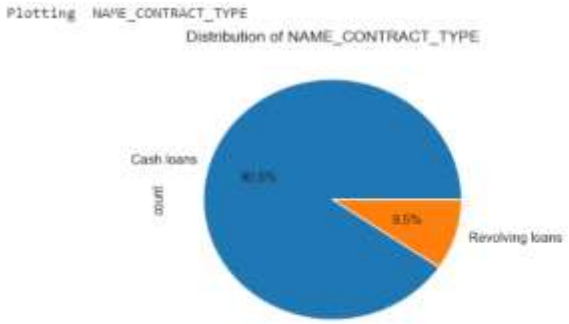
- There are total 49 columns in the application_data.csv with missing values more than 40%
- In the 49 columns, almost every column is either mean, median or mode
- There is no use of these columns.
- These Columns are dropped from the DataFrame.

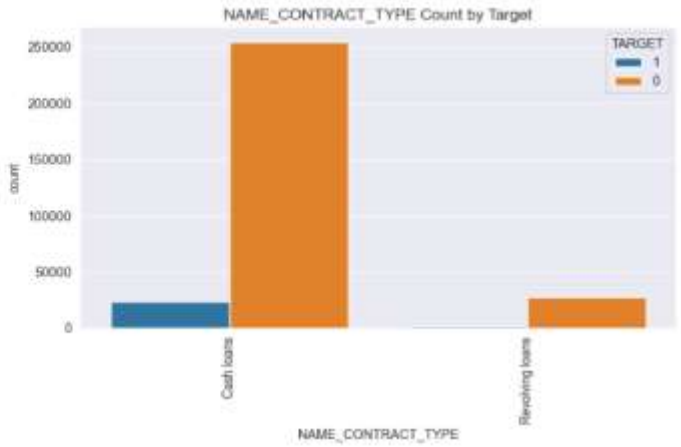
Data Imbalance Analysis

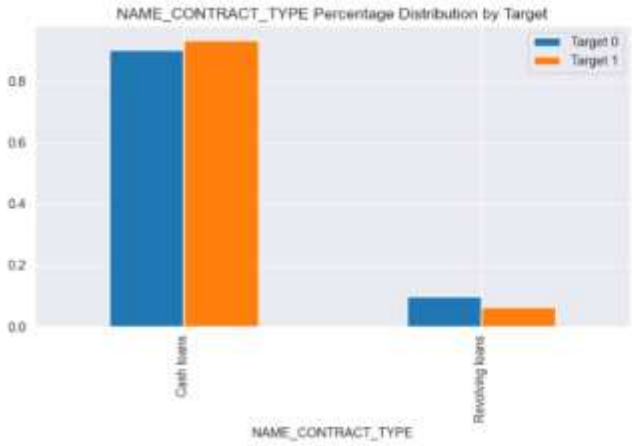


- As the ratio between the values with "TARGET" equals to 0 to the values with "TARGET" equals to 1 is big.
- There is too much data imbalance.
- Due to data imbalance, we have separated the data with "TARGET" equals to 0 and data with "TARGET" equals to 1 in two separate DataFrames.

Univariate Analysis of the Categorical Data



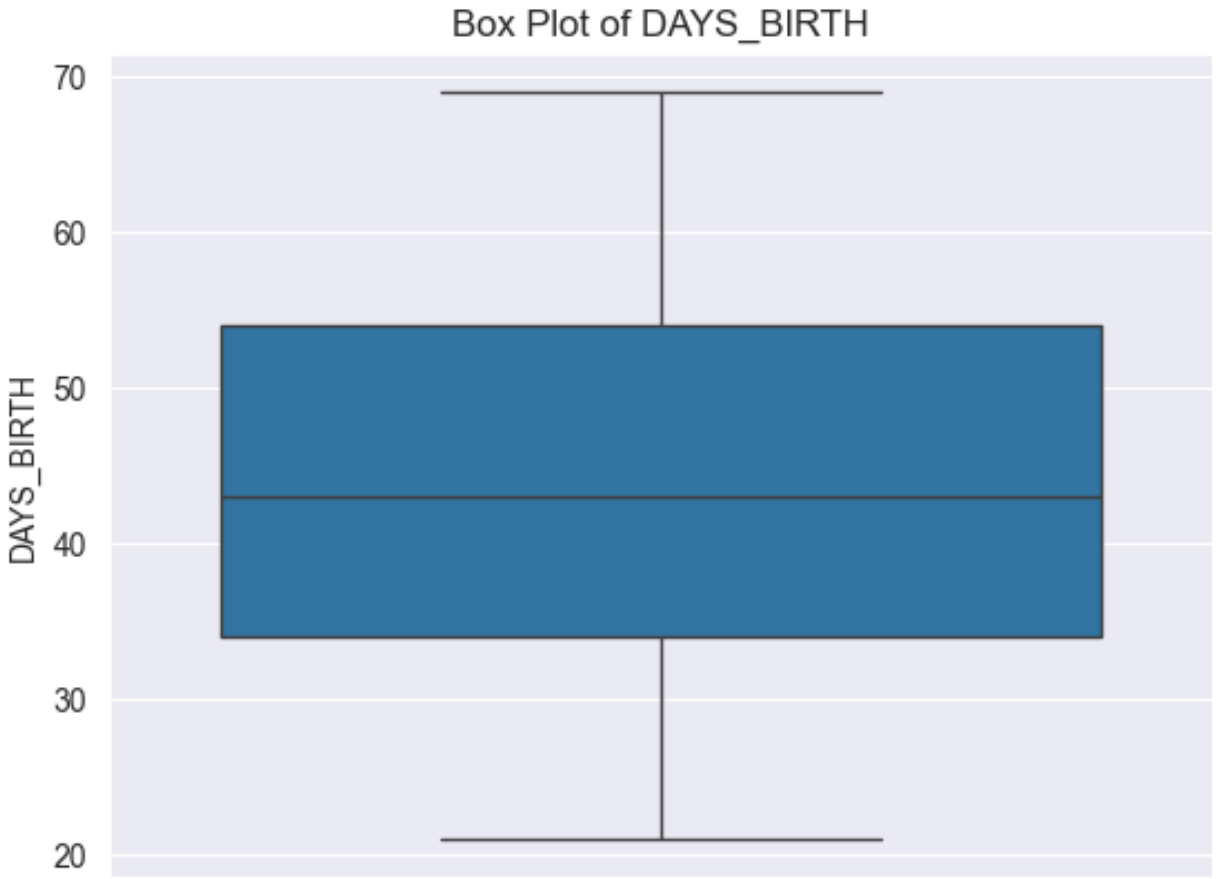




- **Cash loans** have a relatively high percentage of both non-defaulted and defaulted loans, showing some risk associated with this type.
- **Revolving loans**:
  - Show a smaller percentage of defaults in comparison to non-defaults.
  - Likely indicate lower default risk compared to Cash loans, though the sample size is smaller.

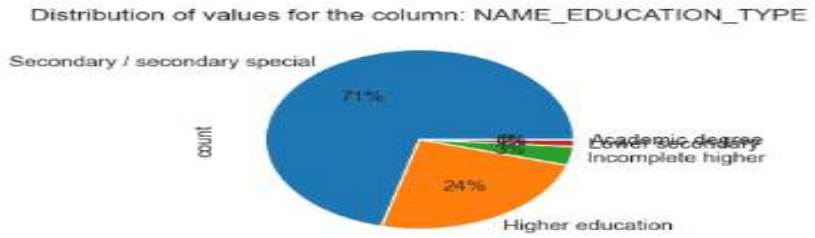
- **For Cash Loan**
  - A large proportion of loans did not default
  - A smaller but noticeable segment did not default
- **For Revolving loan**
  - The majority of contracts did not default
  - Some proportion does default.

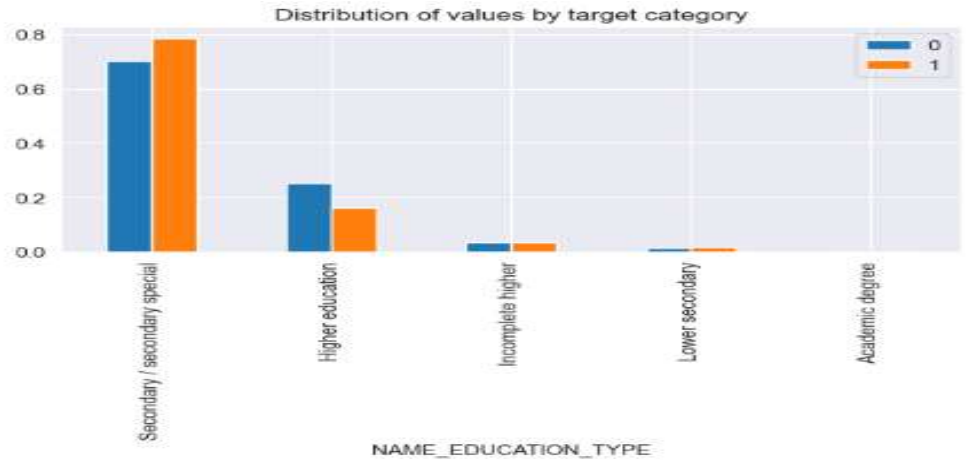
Outlier Analysis

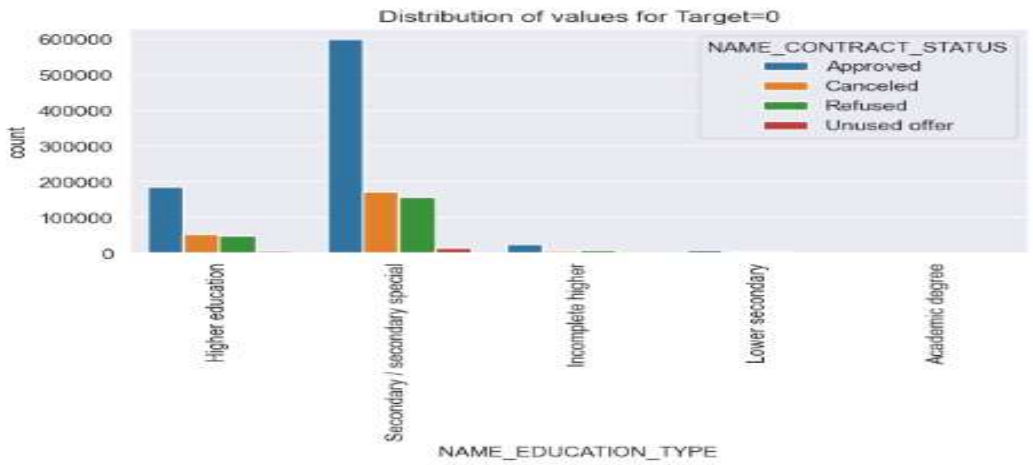


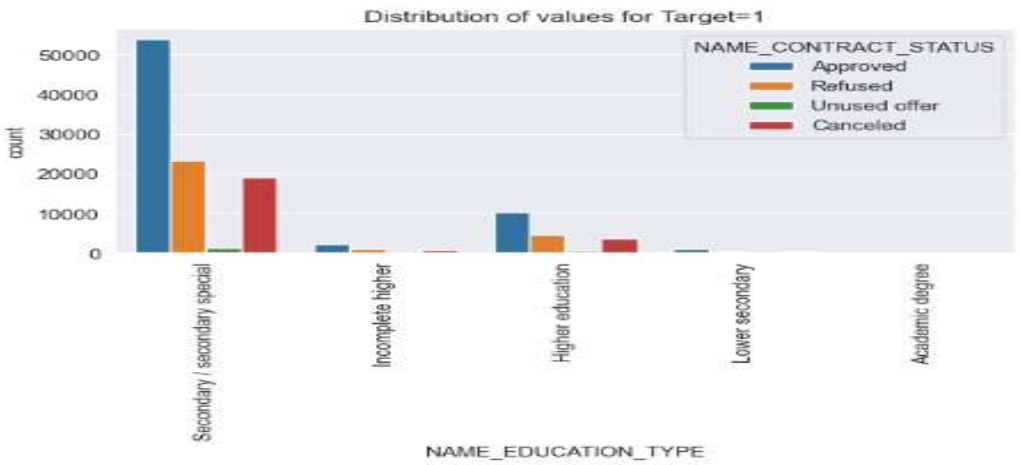
- The Median Age is approximately 40-50 years. This suggests that half of the dataset consists of individuals who are at least 40 years old.
- The Interquartile range (IQR), represented by the box, indicates that the entire age distribution spans roughly from 20 years (lower whisker) to 70 years (upper whisker), indicating the general age range in the dataset.
- There are no points outside the whiskers, suggesting that there are not extreme outliers in terms of age in this dataset.
- The box plot appears relatively symmetrical around the median, suggesting that the age distribution is fairly balanced without strong skewness toward younger or older ages.
- This age range could be relevant for understanding loan risk, as age often correlates with factors like income stability and financial responsibility.

Segmented Univariate Analysis



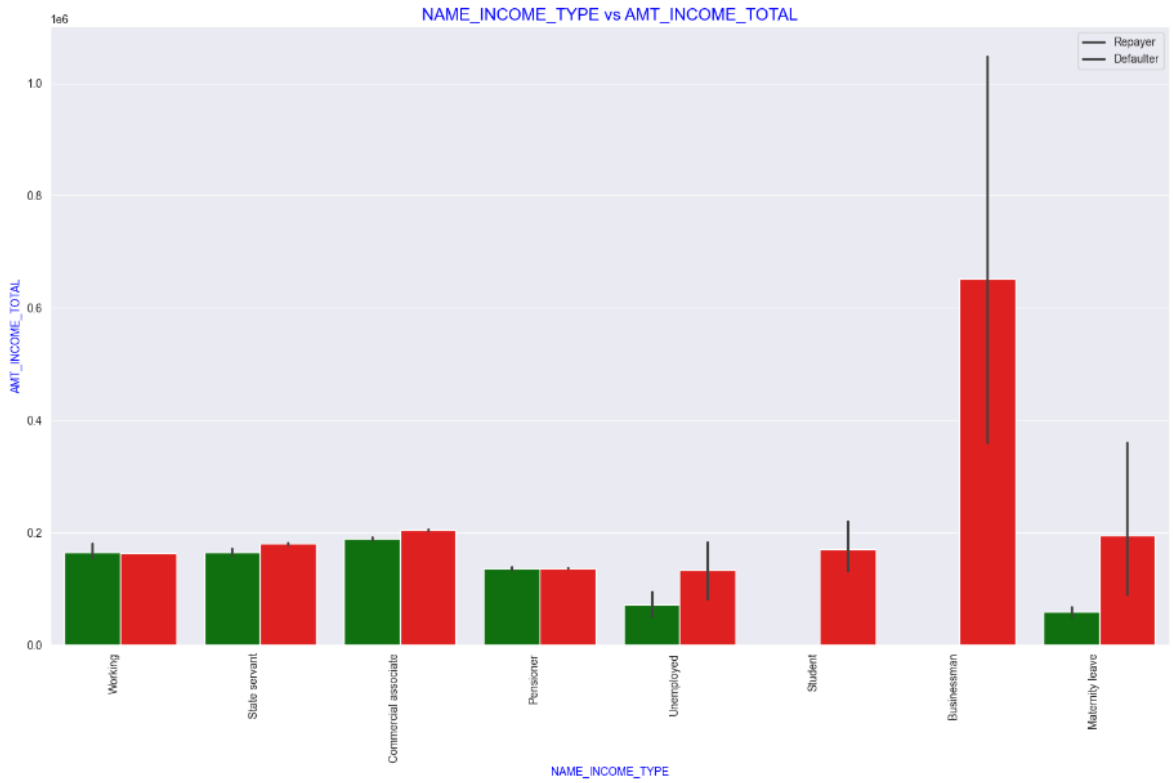






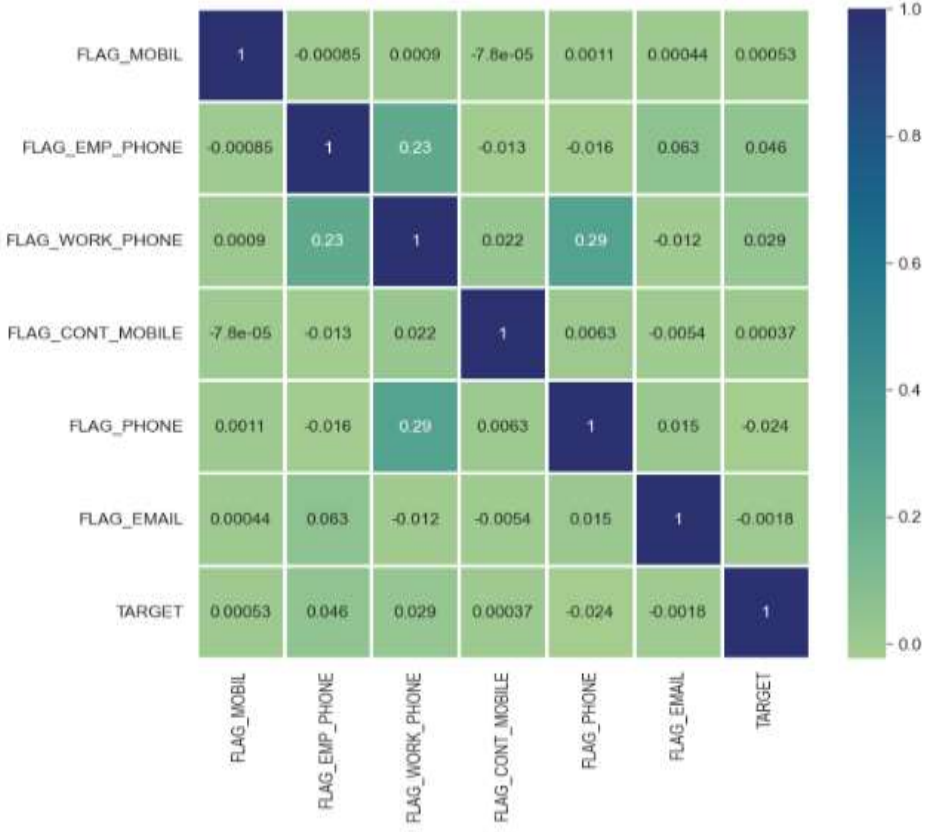
- **Education Distribution**: The majority of applicants (71%) have a **Secondary/Secondary special** education, followed by **Higher education** at 24%. Lower education levels make up a minor portion of the dataset.
- **Default Rates by Education**: Applicants with **Secondary education** have a noticeably higher rate of defaults compared to those with **Higher education**, suggesting that lower education levels may be linked to higher credit risk.
- **Approval vs. Refusal by Education**:
  - **Higher education** applicants have higher approval rates and fewer refusals, indicating a potentially stronger creditworthiness.
  - **Secondary education** applicants show more refusals, aligning with higher default risks.
- **Education as a Risk Factor**: The data suggests that **Higher education** correlates with lower default rates, making it a useful predictor for assessing loan risk.
- **Implication for Risk Models**: Educational background, especially distinguishing between Secondary and Higher education levels, could enhance the accuracy of credit risk assessments.

Bivariate Analysis



- **Income Type Matters:** Repayers and defaulters show distinct income patterns. Repayers tend to have higher income levels across most types.
- **Stable Income, Lower Risk:** Categories like "Commercial Associate" and "Pensioner" have a higher proportion of repaying individuals, suggesting a more stable income source.
- **Unstable Income, Higher Risk:** Categories like "Unemployed" and "Student" have a higher proportion of defaulters, likely due to lower or unstable income levels.
- **More Data Needed:** Categories like "Businessman" and "Maternity Leave" have limited data points, requiring more analysis for conclusive insights.

Correlation Analysis



- Most of the correlations between the variables are weak, indicating a lack of strong linear relationships.
- The "FLAG_WORK_PHONE" variable has a moderate positive correlation with the "TARGET" variable, suggesting that having a work phone might be associated with a higher likelihood of the target outcome.
- The "FLAG_EMAIL" variable has a weak negative correlation with the "TARGET" variable, suggesting that having an email might be associated with a slightly lower likelihood of the target outcome.
- The correlation matrix suggests that there is no strong multicollinearity among the predictor variables, meaning that the variables are not highly correlated with each other. This is good because multicollinearity can make it difficult to interpret the results of a model.

Conclusion

Decisive Factor whether an applicant will be Repayer:

- NAME_EDUCATION_TYPE: Academic degree has less defaults.
- NAME_INCOME_TYPE: Student and Businessmen have no defaults.
- REGION_RATING_CLIENT: RATING 1 is safer.
- ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%
- DAYS_BIRTH: People above age of 50 have low probability of defaulting
- DAYS_EMPLOYED: Clients with 40+ year experience having less than 1% default rate
- AMT_INCOME_TOTAL:Applicant with Income more than 700,000 are less likely to default
- NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, Buying garage are being repayed mostly.
- CNT_CHILDREN: People with zero to two children tend to repay the loans.

Decisive Factor whether an applicant will be Defaulter:

- CODE_GENDER: Men are at relatively higher default rate
- NAME_FAMILY_STATUS : People who have civil marriage or who are single default a lot.
- NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education
- NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot.
- REGION_RATING_CLIENT: People who live in Rating 3 has highest defaults.
- OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as the default rate is huge.
- ORGANIZATION_TYPE: Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed people have relative high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.
- DAYS_BIRTH: Avoid young people who are in age group of 20-40 as they have higher probability of defaulting
- DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.
- CNT_CHILDREN & CNT_FAM_MEMBERS: Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected.
- AMT_GOODS_PRICE: When the credit amount goes beyond 3M, there is an increase in defaulters.

Suggestions

- 90% of the previously cancelled client have actually repayed the loan. Record the reason for cancellation which might help the bank to determine and negotiate terms with these repaying customers in future for increase business opportunity.
- 88% of the clients who were refused by bank for loan earlier have now turned into a repaying client. Hence documenting the reason for rejection could mitigate the business loss and these clients could be contacted for further loans.