Bank Loan Case Study
Final Project-2

Kiruba Shankar S

Data Analytics Trainee

Description:

Perform Exploratory Data Analysis on a loan application dataset to identify missing data, outliers, and data imbalance. Conduct univariate, segmented univariate, and bivariate analyses to uncover factors influencing loan defaults. Identify top correlations for different customer scenarios to enhance loan approval decisions and mitigate financial risks.

Approach:

My approach involved a systematic Exploratory Data Analysis (EDA) using Excel. I addressed missing data with functions like COUNT and AVERAGE, and identified outliers using QUARTILE and IQR, visualizing them with box plots. I assessed data imbalance with COUNTIF and visualized it using pie charts. I conducted univariate, segmented univariate, and bivariate analyses using descriptive statistics and pivot tables, creating histograms and scatter plots. Finally, I calculated correlations using the CORREL function and highlighted top indicators of loan defaults with heatmaps, ensuring comprehensive insights for better loan approval decisions.

Tech-Stack Used:

I used Microsoft Excel for data analysis, also enabled the data analysis add-in to make it less time consuming.

Files:

1. application_data_main: Which contains the Data Cleaning steps, outliers and pivot analysis, Univariate, Segmented Univariate, and Bivariate Analysis, Analysis of Data Imbalance and Correlations.
2. Previous_application_main: Which contains cleaned data.

Business Objectives:

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments. This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants. The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.

A. Identify Missing Data and Deal with it Appropriately:

Application_data file:

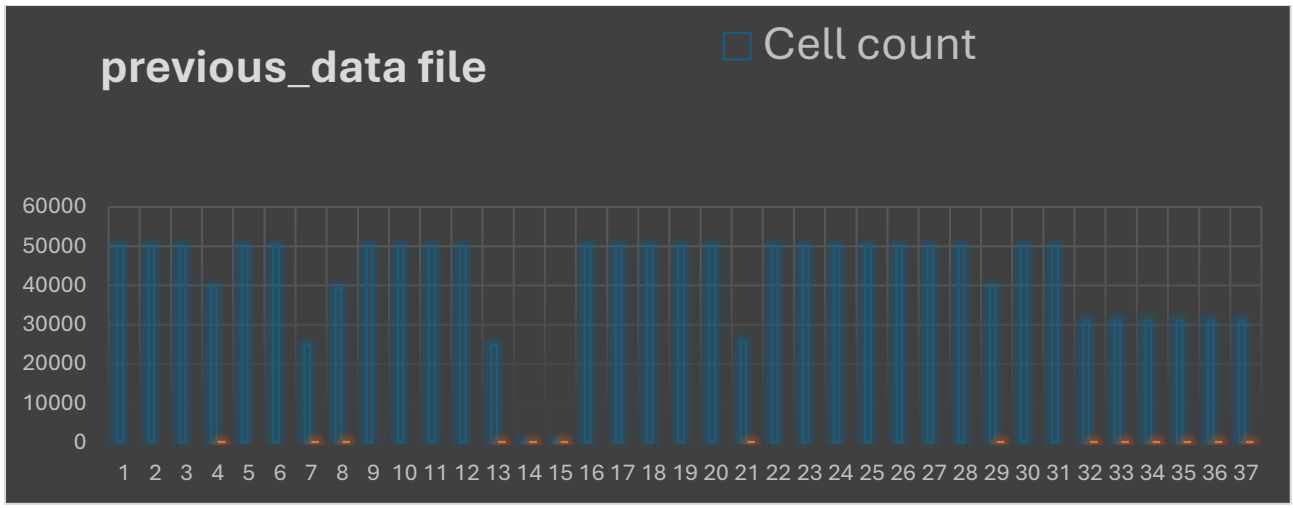
OWN_CAR_AGE, OCCUPATION_TYPE, EXT_SOURCE_1, APARTMENTS_AVG, BASEMENTAREA_AVG, YEARS_BEGINEXPLUATATION_AVG, YEARS_BUILD_AVG, COMMONAREA_AVG, ELEVATORS_AVG, ENTRANCES_AVG, FLOORSMAX_AVG, FLOORSMIN_AVG, LANDAREA_AVG, LIVINGAPARTMENTS_AVG, LIVINGAREA_AVG, NONLIVINGAPARTMENTS_AVG, NONLIVINGAREA_AVG, APARTMENTS_MODE, BASEMENTAREA_MODE, YEARS_BEGINEXPLUATATION_MODE, YEARS_BUILD_MODE, COMMONAREA_MODE, ELEVATORS_MODE, ENTRANCES_MODE, FLOORSMAX_MODE, FLOORSMIN_MODE, LANDAREA_MODE, LIVINGAPARTMENTS_MODE, LIVINGAREA_MODE, NONLIVINGAPARTMENTS_MODE, NONLIVINGAREA_MODE, APARTMENTS_MEDI, BASEMENTAREA_MEDI, YEARS_BEGINEXPLUATATION_MEDI, YEARS_BUILD_MEDI, COMMONAREA_MEDI, ELEVATORS_MEDI, ENTRANCES_MEDI, FLOORSMAX_MEDI, FLOORSMIN_MEDI, LANDAREA_MEDI, LIVINGAPARTMENTS_MEDI, LIVINGAREA_MEDI, NONLIVINGAPARTMENTS_MEDI, NONLIVINGAREA_MEDI, FONDKAPREMONT_MODE, HOUSETYPE_MODE, TOTALAREA_MODE, WALLSMATERIAL_MODE, EMERGENCYSTATE_MODE.

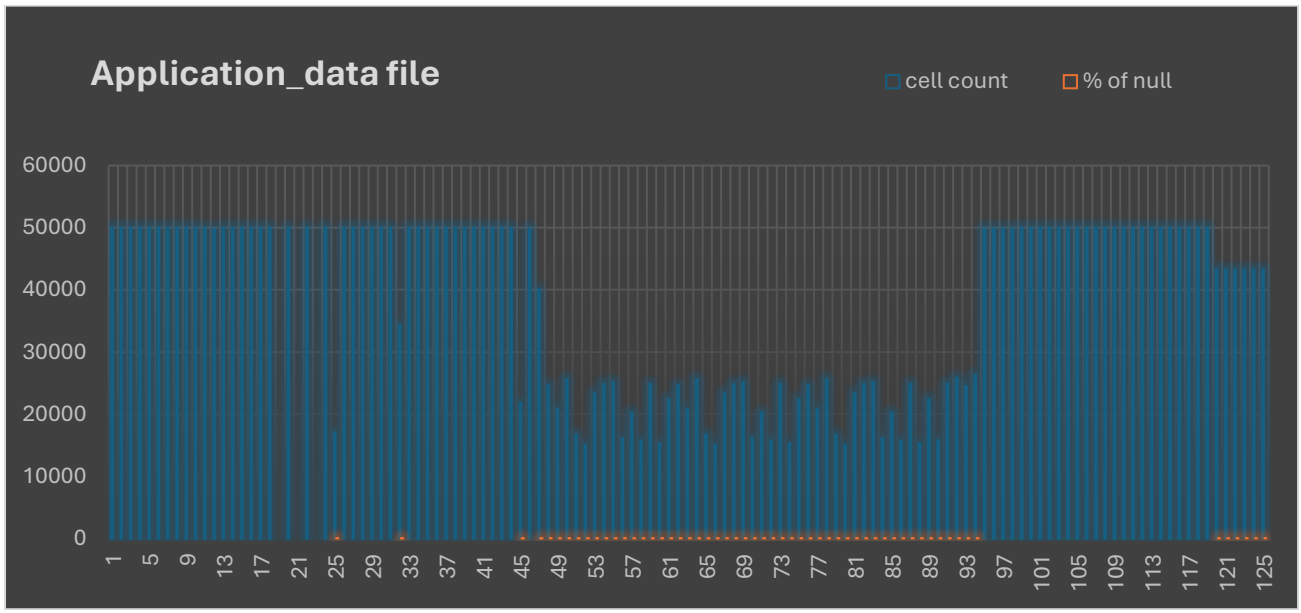
previous_data file:

AMT_DOWN_PAYMENT, RATE_DOWN_PAYMENT, RATE_INTEREST_PRIMARY, RATE_INTEREST_PRIVILEGED, NAME_TYPE_SUITE, DAYS_FIRST_DRAWING, DAYS_FIRST_DUE, DAYS_LAST_DUE_1ST_VERSION, DAYS_LAST_DUE, DAYS_TERMINATION, NFLAG_INSURED_ON_APPROVAL

Cleared 50 columns from applications_data and 10 columns from previous_data, hence cleared the column which has above 30 percentage of the null values, by using CountA function to identify the number of rows used and by using the formula (=1-B1/B1) to calculate the percentage of the null values.

To determine which parts of the dataset are missing, used Excel's built-in features and functions to address the issue. When it comes to data imputation, we have used methods like AVERAGE or MEDIAN in addition to Excel functions like COUNT, ISBLANK, and IF to identify any missing information. In order to visually represent the distribution of missing values across variables used a bar chart.





B. **Identify Outliers in the Dataset**:

Outliers can significantly impact the analysis and distort the results.

- To find and describe outliers in the dataset using Excel's statistical tools, paying special attention to numerical variables.
- To identify any outliers in the dataset, use Excel tools like QUARTILE, IQR and conditional formatting. To determine if these outliers are valid data items or require more investigation, use thresholds or business rules.
- To highlight the existence of outliers and graphically depict the distribution of numerical variables in the dataset, create box or scatter plots.

<u>Outliers:</u>

- The mean income is around 170,767, with a large standard deviation of 531,819, indicating high variability in income.
- The data has a high range, from 25,650 to 117,000,000, and is positively skewed with a skewness of 212.08.
- Significant outliers in high-income values might skew the analysis and should be further investigated or treated.

QUARTILE 1:	112500
QUARTILE 2:	145800
QUARTILE 3:	202500

Inter quartile	90000
upper limit	337500
Lower Limit	-22500

Quartiles and Interquartile Range (IQR):

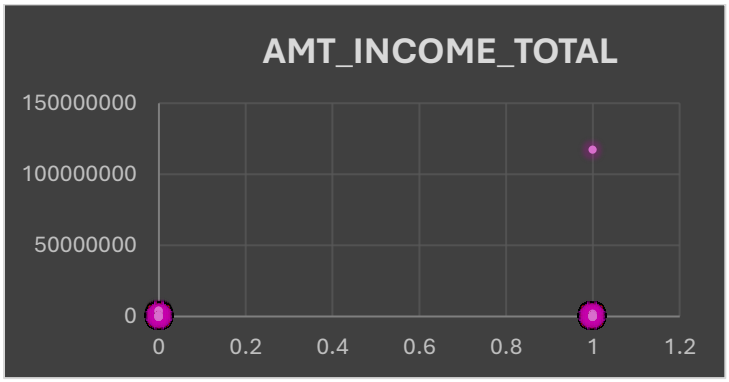
Quartile 1 (Q1): 112,500
Quartile 2 (Q2): 145,800
Quartile 3 (Q3): 202,500

Inter quartile (IQR): Q3 - Q1 = 90,000
Upper Limit: Q3 + 1.5 * IQR = 337,500
Lower Limit: Q1 - 1.5 * IQR = -22,500

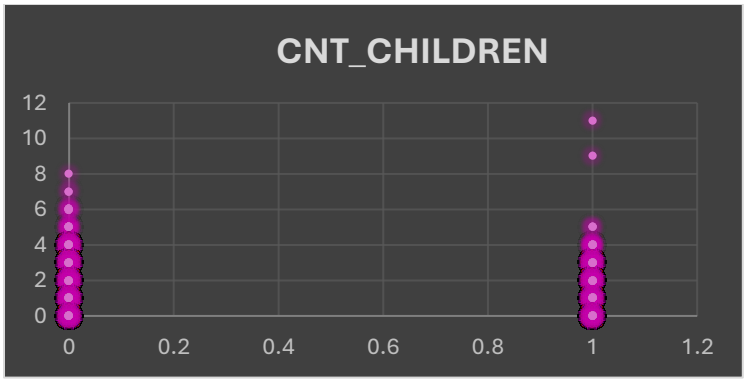
Insights:

- The mean income is around 170,767, with a large standard deviation of 531,819, indicating high variability in income.
- The data has a high range, from 25,650 to 117,000,000, and is positively skewed with a skewness of 212.08.
- Significant outliers in high-income values might skew the analysis and should be further investigated or treated.
- There is a significant imbalance in the dataset, with the majority having no children. This imbalance should be addressed during analysis and model building to ensure that predictive models are not biased.
- The number of children can significantly impact an applicant's financial situation and their ability to repay loans. It is crucial to analyze this variable in conjunction with income and other financial metrics to make informed lending decisions.
- High prevalence of 0 days suggests potential data issues or new applicants; ensure data accuracy and correlate with other metrics for a comprehensive loan risk assessment.

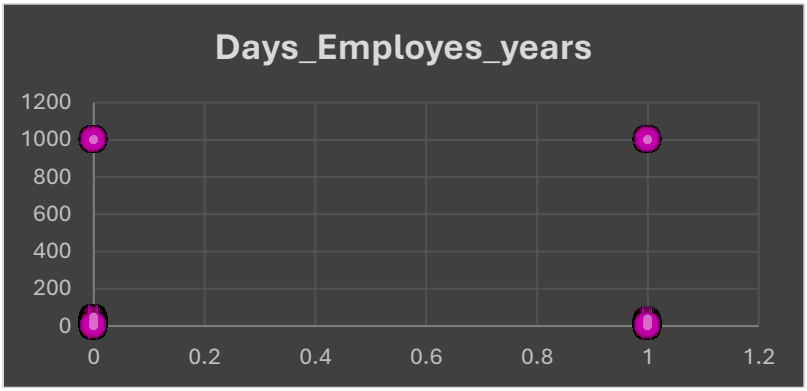
Outlier 1:



Outlier 2:



Outlier 3:



C. Analyze Data Imbalance:

Data Distribution:

- Class 0 (Non-Default): 45,973 instances (92% of the dataset).
- Class 1 (Default): 4,026 instances (8% of the dataset).

Value	Count of Target
1	4026
0	45973
Ratio	11.41902633

Imbalance Ratio:

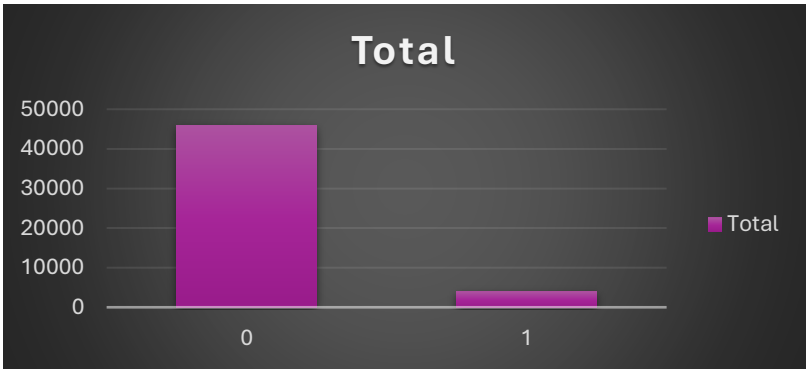
- The ratio of non-default to default cases is approximately 11.42:1, indicating a significant imbalance where non-default cases are substantially more prevalent.

Insights:

- The high imbalance may lead to biased models that favor the majority class (non-default). This could affect the model's ability to accurately predict the minority class (default).

- The dataset shows significant class imbalance, with non-default cases vastly outnumbering default cases. This imbalance needs to be addressed to ensure accurate and reliable predictions in your model.

Row Labels	Sum of Count of Target
0	45973
1	4026
Grand Total	**49999**



D. Perform Univariate, Segmented Univariate, and Bivariate Analysis:

For gaining insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

1. Segmented Univariate Analysis:

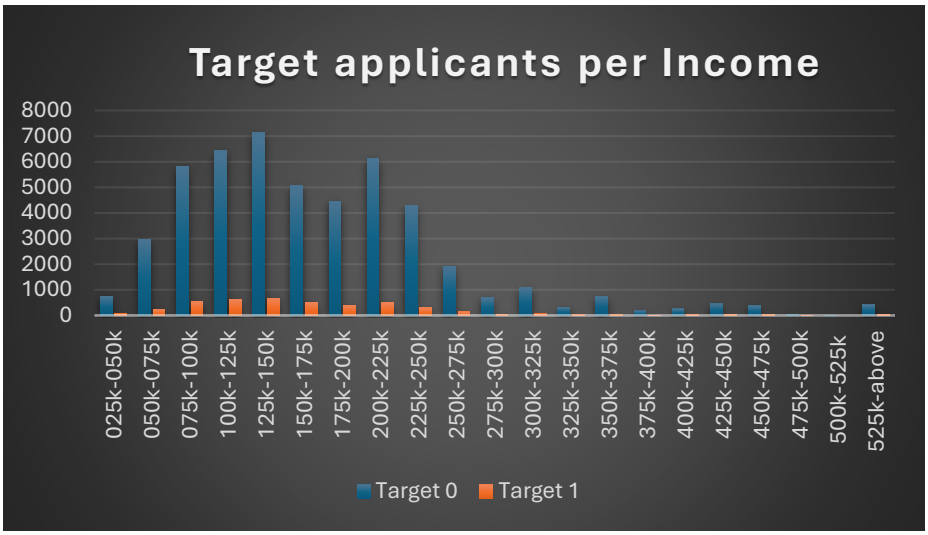
- Higher Default Rates for Smaller Loans: Default rates are higher in lower loan ranges (e.g., 25k-50k), with smaller amounts showing increased default percentages.

- Increasing Defaults with Loan Size: The number of defaults rises with loan amounts, but the default rate per amount tends to decrease as loan sizes grow.

- Low Default Rates for Very High Loans: Extremely high loan amounts (500k+) exhibit very low default rates, suggesting financial stability or stringent lending criteria.

- Significant Drop in Defaults for Extreme Amounts: Default rates drop significantly for the highest loan ranges, indicating minimal risk among very high loan borrowers.

smaller loan amounts are associated with higher default rates, while larger amounts show decreasing default rates relative to the number of applicants. Extreme high loan amounts have minimal default occurrences, indicating potential stability among high-value borrowers or strict lending standards.



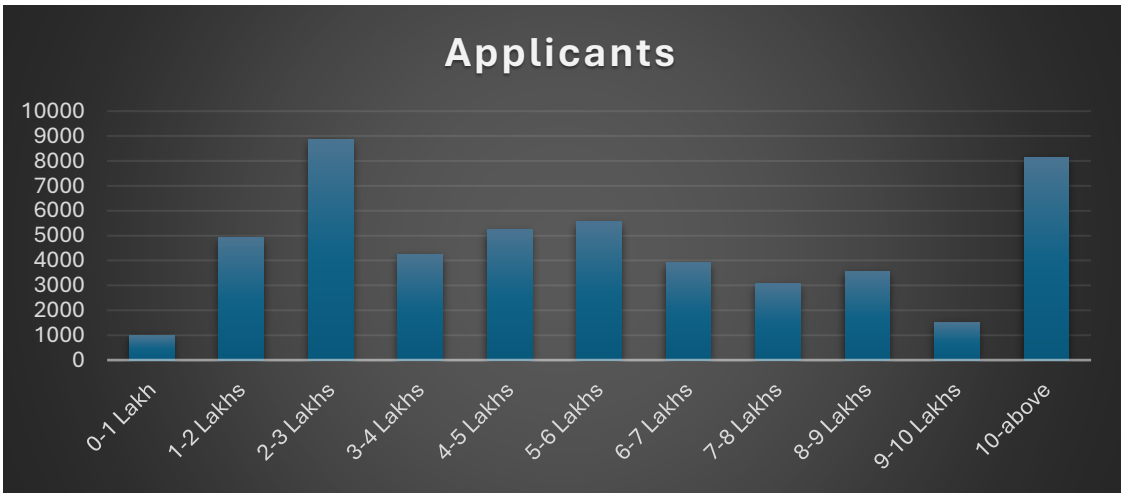
2. <u>Univariate Analysis:</u>

- Middle Range Popularity: The 2-3 Lakhs (8,849 applicants) and 5-6 Lakhs (5,554 applicants) credit bins have the highest applicant concentrations.

- High Demand for Large Loans: The 10 Lakhs and above bin has a substantial number of applicants (8,146), indicating strong demand for large credit amounts.

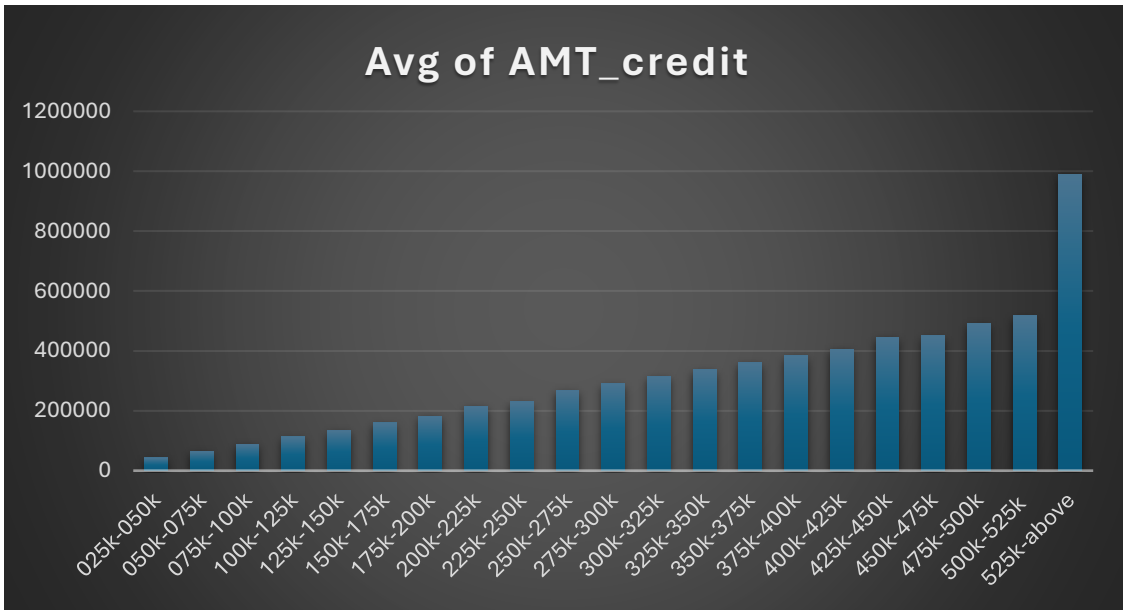
- Least Popular Bins: The 0-1 Lakh (989 applicants) and 9-10 Lakhs (1,524 applicants) bins have the fewest applicants, suggesting lower demand for these credit ranges.

- Stable Mid-Range Distribution: Credit bins from 3-7 Lakhs have a balanced distribution, each with 3,000 to 5,000+ applicants, indicating stable demand in these ranges.



3. Bivariate Analysis

- Positive Correlation: Higher income bins correspond to higher average credit amounts, with a clear upward trend from ₹43,179 (025k-050k) to ₹517,000 (500k-525k).
- Significant Growth: Average credit amounts increase significantly from ₹65,308 (050k-075k) to ₹159,260 (150k-175k) as income rises.
- Exponential Increase in High-Income Bins: Average credit amounts rise sharply in high-income bins, reaching ₹517,000 (500k-525k) and peaking at ₹989,409 (525k-above).
- Steady Growth in Mid-Income Ranges: Mid-income bins (100k-325k) show consistent growth in average credit amounts, reflecting a balanced increase with income.



There is a positive correlation between income and average credit amounts, with higher income bins corresponding to higher average credit amounts, ranging from ₹43,179 in the 025k-050k bin to ₹517,000 in the 500k-525k bin. The data shows significant growth in average credit amounts as income rises, notably increasing from ₹65,308 in the 050k-075k bin to ₹159,260 in the 150k-175k bin. This trend becomes even more pronounced in high-income bins, where average credit amounts rise sharply, reaching ₹517,000 in the 500k-525k bin and peaking at ₹989,409 in the 525k-above bin. Mid-income bins (100k-325k) display steady growth in average credit amounts, indicating a balanced increase with income.

E. Identify Top Correlations for Different Scenarios:

By understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

1. Applicants who made payment ON TIME:

Target 0	CNT_CHILDREN	NCOME_TOTAL	AMT_CREDIT	ON_RELATIVE	_Birth_Years	yes_years	l_publish_Years	ING_CLIENT
CNT_CHILDREN	1	0.036319722	0.005705458	-0.024912809	-0.33587627	-0.245522	0.032537221	0.02132465
AMT_INCOME_TOTAL	0.036319722	1	0.377965752	0.181941261	-0.07376942	-0.161681	-0.032286356	-0.20501152
AMT_CREDIT	0.005705458	0.377965752	1	0.095539444	0.051084182	-0.074733	0.008290189	-0.10251819
REGION_POPULATION_RELATIVE	-0.024912809	0.181941261	0.095539444	1	0.030435419	-0.006767	0.002236288	-0.53936191
Days_Birth_Years	-0.335876269	-0.073769425	0.051084182	0.030435419	1	0.6234747	0.270073313	-0.00895244
Days_Employes_years	-0.245521512	-0.161680938	-0.074733443	-0.006767142	0.623474675	1	0.274516224	0.04095027
Days_Id_publish_Years	0.032537221	-0.032286356	0.008290189	0.002236288	0.270073313	0.2745162	1	0.0081684
REGION_RATING_CLIENT	0.021324653	-0.205011517	-0.102518185	-0.539361913	-0.00895244	0.0409503	0.008168404	1

- Negative Correlation with Age and Employment Duration: The number of children (CNT_CHILDREN) shows a moderate negative correlation with age (Days_Birth_Years) and employment duration (Days_Employes_years), indicating younger applicants or those with less employment history are more likely to have children.

- Positive Correlation between Income and Credit Amount: There is a strong positive correlation between total income (AMT_INCOME_TOTAL) and credit amount (AMT_CREDIT), suggesting that applicants with higher incomes tend to secure higher credit amounts.

- Region Population and Credit Ratings: REGION_POPULATION_RELATIVE shows a strong negative correlation with REGION_RATING_CLIENT, implying that applicants from more populated regions have lower client ratings.

- Age and Employment Duration: A strong positive correlation exists between age (Days_Birth_Years) and employment duration (Days_Employes_years), indicating that older applicants typically have longer employment histories.

2. Applicants who have payment DIFFICULTIES:

Target 1	CNT_CHILDREN	NCOME_TOTAL	AMT_CREDIT	ON_RELATIVE	_Birth_Years	oyes_years	l_publish_Years	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.010110177	0.007601905	-0.020359154	-0.2496732	-0.189773	0.042360717	0.055515557
AMT_INCOME_TOTAL	0.010110177	1	0.015271444	-0.006180303	-0.00903366	-0.011759	0.009122006	-0.012846697
AMT_CREDIT	0.007601905	0.015271444	1	0.067775624	0.142506035	0.0187822	0.043771901	-0.045024534
REGION_POPULATION_RELATIVE	-0.020359154	-0.006180303	0.067775624	1	0.016468731	0.0077101	0.005118563	-0.430032303
Days_Birth_Years	-0.2496732	-0.009033662	0.142506035	0.016468731	1	0.5882428	0.247896571	-0.045027112
Days_Employes_years	-0.189773227	-0.011758681	0.018782223	0.007710059	0.588242824	1	0.232661912	-0.009237108
Days_Id_publish_Years	0.042360717	0.009122006	0.043771901	0.005118563	0.247896571	0.2326619	1	-0.025335227
REGION_RATING_CLIENT	0.055515557	-0.012846697	-0.045024534	-0.430032303	-0.04502711	-0.009237	-0.025335227	1

- Income and Credit Amount Relationship: There is a weak positive correlation between total income (AMT_INCOME_TOTAL) and credit amount (AMT_CREDIT), indicating a less pronounced relationship compared to those who pay on time.
- Credit Amount and Age: The correlation between credit amount (AMT_CREDIT) and age (Days_Birth_Years) is stronger, suggesting that older applicants are more likely to have higher credit amounts.
- Region Population and Credit Ratings: Similar to applicants who pay on time, REGION_POPULATION_RELATIVE shows a strong negative correlation with REGION_RATING_CLIENT, indicating lower client ratings in more populated regions.
- Age and Employment Duration: The positive correlation between age (Days_Birth_Years) and employment duration (Days_Employes_years) is evident, highlighting that older applicants typically have longer employment histories, similar to those who pay on time.

The income and credit amount correlation are more pronounced in applicants who pay on time, while the correlation between credit amount and age is stronger in those with payment difficulties. These insights help in understanding the factors influencing payment behavior and creditworthiness.

Conclusion:

The loan application data analysis reveals significant data imbalance, with 92% non-defaults and 8% defaults. Most applicants fall into middle-income brackets, impacting risk profiles. Higher income correlates with increased credit amounts. Applicants with payment difficulties show distinct correlation patterns, such as stronger ties between credit amount and age. Key indicators like age, employment history, income, and regional population are crucial in predicting loan defaults. These insights aid in refining credit risk models for better lending decisions.