Bank Loan Case Study

1. Overview:

In the Bank Loan Case Study, our primary goal is to perform Exploratory Data Analysis (EDA) on a vast dataset of loan applications. We work for a finance company specializing in lending various types of loans to urban customers. The company faces a significant challenge: identifying capable applicants and ensuring that they are not rejected. This project is driven by the need to minimize the risks associated with loan defaults.

The company encounters two key risks when processing loan applications:

1. Risk of Rejecting Capable Applicants: If a qualified applicant is denied a loan, the company loses business opportunities.
2. Risk of Loan Defaults: If an applicant who cannot repay the loan is approved, the company faces financial losses.

The dataset we are analyzing contains information about loan applications and includes two distinct scenarios:

Customers with Payment Difficulties: These are customers who experienced late payments of more than X days on at least one of the first Y installments of their loans.

All Other Cases: These are scenarios where payments were made on time.

When a customer applies for a loan, there are four potential outcomes:

1. Approved: The company has approved the loan application.
2. Cancelled: The customer cancelled the application during the approval process.
3. Refused: The company rejected the loan application.
4. Unused Offer: The loan was approved, but the customer did not utilize it.

Our project's overarching objective is to leverage EDA to comprehend how customer attributes and loan attributes influence the likelihood of loan defaults. The specific business goals include:

- Identifying patterns that indicate whether a customer may have difficulty repaying their installments.
- Using these patterns to make informed decisions, such as denying the loan, reducing the loan amount, or offering loans at higher interest rates to risky applicants.
- Gaining insights into the key factors contributing to loan defaults, enabling the company to make more accurate decisions regarding loan approvals.

To achieve these objectives, we will perform various data analytics tasks:

A. Handling Missing Data: We will identify and address missing data in the loan application dataset using appropriate Excel functions and features.

B. Detecting Outliers: Outliers can impact our analysis, so we will identify and examine outliers in the dataset, focusing on numerical variables.

C. Analyzing Data Imbalance: We will assess data imbalance and calculate the ratio of data imbalance to understand its effect on our analysis, particularly for binary classification problems.

D. Exploratory Data Analysis: This phase involves univariate, segmented univariate, and bivariate analysis to understand the driving factors of loan default.

E. Correlation Analysis: We will segment the dataset based on different scenarios and identify top correlations for each segmented data. This will help us uncover strong indicators of loan default.

2. Tech-Stack Used:

Microsoft Excel

Google Drive for report submission

Excel Hyperlink:

Excel File Link

3. Approach:

The project revolves around utilizing Exploratory Data Analysis (EDA) techniques to mitigate risks in the lending business. The aim is to identify patterns that indicate the likelihood of loan default among customers, helping make informed lending decisions and manage risk effectively.

1. Data Preparation:
    - Obtain and review the two extensive datasets: current and previous loan applications.
    - Identify and eliminate unnecessary columns that won't contribute to risk assessment.
    - Handle missing data appropriately using Excel functions and features.
    - Detect and manage outliers in numerical variables to ensure the data's quality.
2. Data Imbalance Assessment:
    - Analyze the distribution of the target variable to assess data imbalance.
    - Calculate the data imbalance ratio using Excel functions and visualize the class frequencies.
3. Exploratory Data Analysis (EDA):
    - Begin with univariate analysis to understand the distribution of individual variables.
    - Proceed with segmented univariate analysis to compare variable distributions across different scenarios (e.g., customers with payment difficulties vs. others).
    - Conduct bivariate analysis to explore relationships between variables and the target variable using Excel's features.
    - Visualize the data using histograms, bar charts, box plots, stacked bar charts, grouped bar charts, scatter plots, or heatmaps to gain insights into risk factors associated with lending.
4. Correlation Analysis:
    - Segment the dataset based on scenarios, such as clients with payment difficulties and all other cases.
    - Calculate correlation coefficients between variables and the target variable within each segment.
    - Rank the correlations to identify the top indicators of loan default for each scenario.
    - Visualize the correlations through correlation matrices or heatmaps, highlighting the most influential variables.

Dataset Detail:

Application Data (application_data.csv):

- Columns: 127

- Rows: 50,000

Insights:

- Client Assessment: This dataset provides detailed information about loan applicants, including their gender, age, family status, income, and more. It's essential for assessing a client's creditworthiness and the level of risk associated with lending to them.
- Loan Details: Information about the loans applied for, including the credit amount, annuity, goods price, and contract type, helps in understanding the specific financial requirements of clients and the types of loans they are seeking.
- Property and Housing Details: Details about the type of property the client owns (car, realty), housing situation, and property price can offer insights into the client's financial stability and repayment capacity.
- Employment and Income: Information about the client's employment, such as occupation, organization type, and days employed, is useful for evaluating the client's job stability and income level.
- Region and External Sources: Data related to the client's region, population, and external sources can be used to assess the economic conditions of the client's area and their credit history.
- Loan Outcome (TARGET): The target variable categorizes clients into those with payment difficulties and those without. It's crucial for risk assessment and identifying clients more likely to default on their loans.

Previous Application Data (previous_application.csv):

- Columns: 37
- Rows: 50,000

Insights:

- Historical Loan Behavior: This dataset contains information about clients' past loan applications and their outcomes. It offers insights into the client's historical behavior, such as whether they repaid loans on time or had payment difficulties.
- Contract and Product Details: It provides information on the type of contracts and products clients have applied for in the past. This can help in understanding their preferences and risk profiles.
- Payment and Interest Rates: Details about payment methods, down payment rates, and interest rates offer insights into how clients manage their loans and the terms they are comfortable with.
- Client Behavior and History: Information about client behavior, such as the purpose of the loan and the reason for any rejections, can provide additional context for risk assessment.
- Timeline and Insurance: The dataset includes timelines related to loan disbursements and insurance requests, which are valuable for understanding the client's financial planning and risk mitigation.

Usefulness:

- These datasets are valuable for conducting risk assessments in the banking and financial services industry. They enable data analysts to identify patterns, correlations, and risk factors associated with loan defaults.
- The data can be used to build predictive models that assess the likelihood of loan default, helping financial institutions make informed lending decisions and manage risk effectively.
- Understanding client behavior, historical loan performance, and financial attributes aids in making better decisions about loan approval, ultimately reducing potential financial losses.
- These datasets are a practical demonstration of how data can be utilized to mitigate risks in the lending business, making it a valuable resource for risk analytics in the banking sector.

4. Insights:

Task 1: Identify Missing Data and Deal with it Appropriately:

In this task, I conducted a comprehensive analysis of the dataset to identify the columns with missing values. Here are the key insights:

- SK_ID_CURR and TARGET: These columns have no missing values, indicating that every record has a unique client ID, and the target variable is well-defined.
- Numerical Columns: Most numerical columns, such as CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT, and more, have no missing values. This suggests that essential financial and demographic information is complete for all clients.
- AMT_ANNUITY and AMT_GOODS_PRICE: These columns have a small percentage of missing values (0.002% and 0.076%, respectively). Imputing these missing values with the median or mode is a reasonable approach.
- Categorical Columns: Columns like NAME_TYPE_SUITE, OCCUPATION_TYPE, and FONDKAPREMONT_MODE have a relatively high percentage of missing values, ranging from 31.31% to 69.92%. Given the substantial amount of missing data, these columns may not be suitable for direct analysis.
- Other Numerical Columns: Several numerical columns related to properties, education, and more have a high percentage of missing values, ranging from 50.15% to 69.92%. Imputing these columns may lead to biased results.
- FLAG_DOCUMENT_X and DAYS_X: These columns indicate binary flags and days-related variables. They have no missing values, ensuring that these fields are well-populated.
- AMT_REQ_CREDIT_BUREAU_X: These columns have a moderate percentage of missing values (13.47%). Imputing these missing values or using a specific strategy for handling credit bureau requests is advisable.

Data Handling Approach: For columns with more than 30% missing values (e.g., FONDKAPREMONT_MODE, COMMONAREA_AVG), we recommend removing them from the dataset as they might not provide significant insights due to the high missing data.

For columns with a small percentage of missing values (e.g., AMT_ANNUITY, AMT_GOODS_PRICE), imputing the missing values using the median or mode is a reasonable approach to retain valuable information.

For columns with a substantial percentage of missing values (e.g., NAME_TYPE_SUITE, OCCUPATION_TYPE), it's crucial to carefully consider whether imputation is suitable or if alternative strategies should be employed to address the missing data.

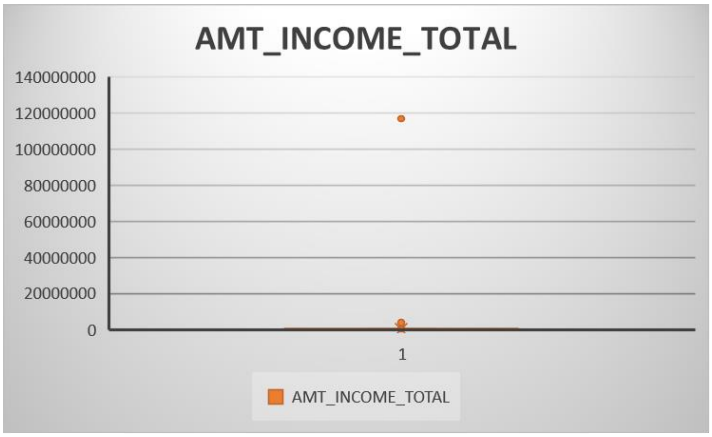
The data cleansing process should ensure that the dataset is as complete as possible while preserving the integrity of the analysis. This will enable more accurate insights and modeling in subsequent tasks.

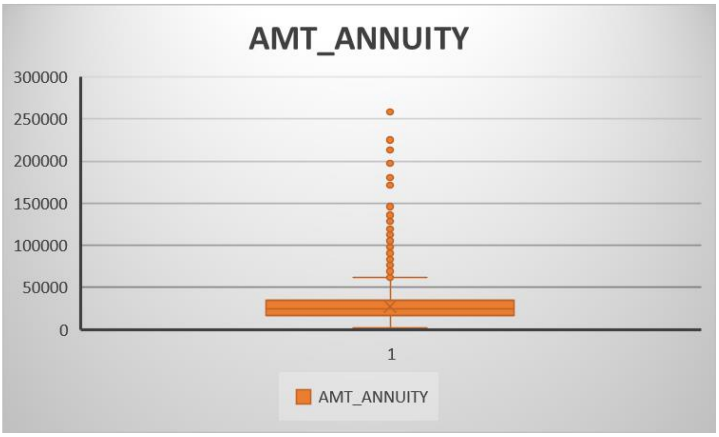
Task 2: Outlier Detection:

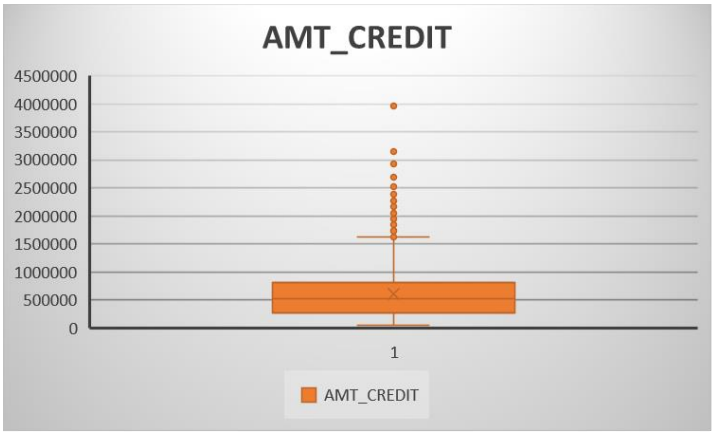
In this task, I performed an analysis to identify outliers in the loan application dataset. The following are the key insights and findings from this task:

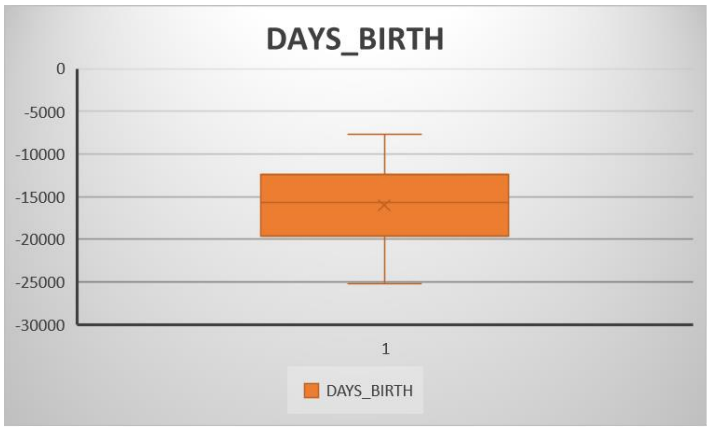
Descriptive Statistics:

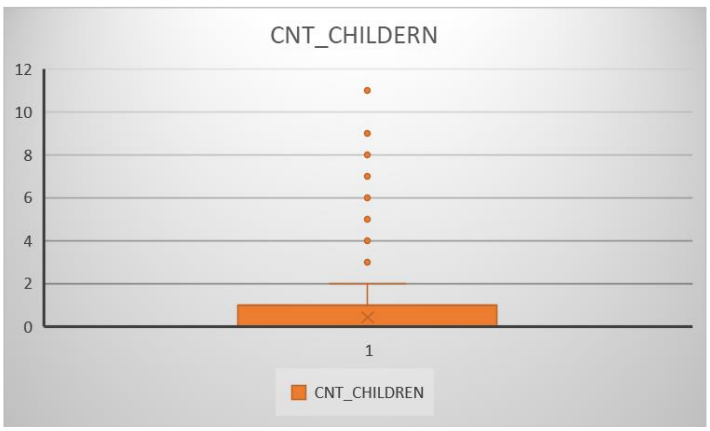
- We calculated descriptive statistics for numerical variables, including the mean, standard error, median, mode, standard deviation, sample variance, kurtosis, skewness, range, minimum, maximum, sum, and count for each variable.

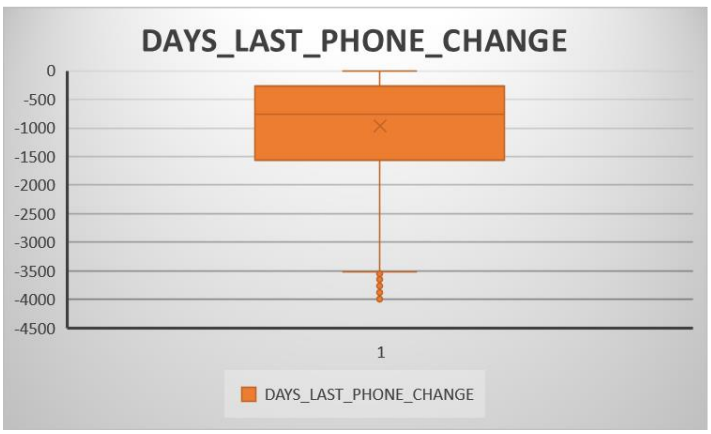










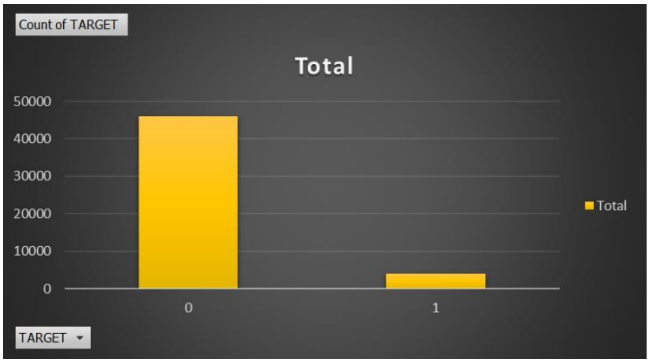


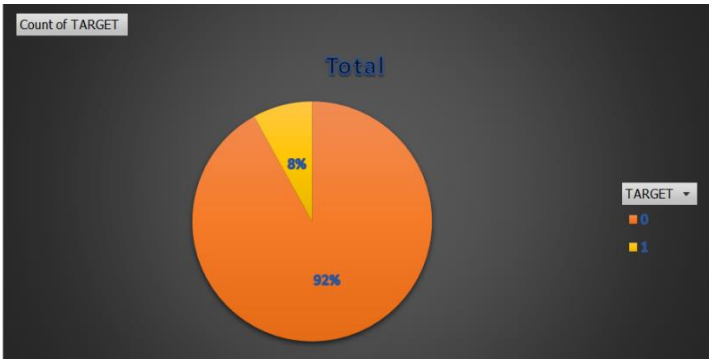
Outliers Detection:

- To identify outliers, we focused on the following numerical variables: CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, DAYS_BIRTH, DAYS_EMPLOYED, and DAYS_LAST_PHONE_CHANGE.

- We used the Interquartile Range (IQR) method to detect outliers. The IQR is a measure of statistical dispersion and is calculated as the difference between the first quartile (Q1) and the third quartile (Q3). Outliers are typically defined as values outside the range (Q1 - 1.5 * IQR) to (Q3 + 1.5 * IQR).
- For each variable, we calculated the quartiles (Q1 and Q3), the IQR, and defined upper and lower limits for identifying outliers.

Task 3: Data Imbalance:





Data Imbalance Analysis:

- We examined the distribution of the target variable "TARGET," which is a binary variable indicating loan outcomes.
- We found that there are 45,973 instances (approximately 92%) with a target value of 0 and 4,026 instances (approximately 8%) with a target value of 1.

Data Imbalance Ratio:

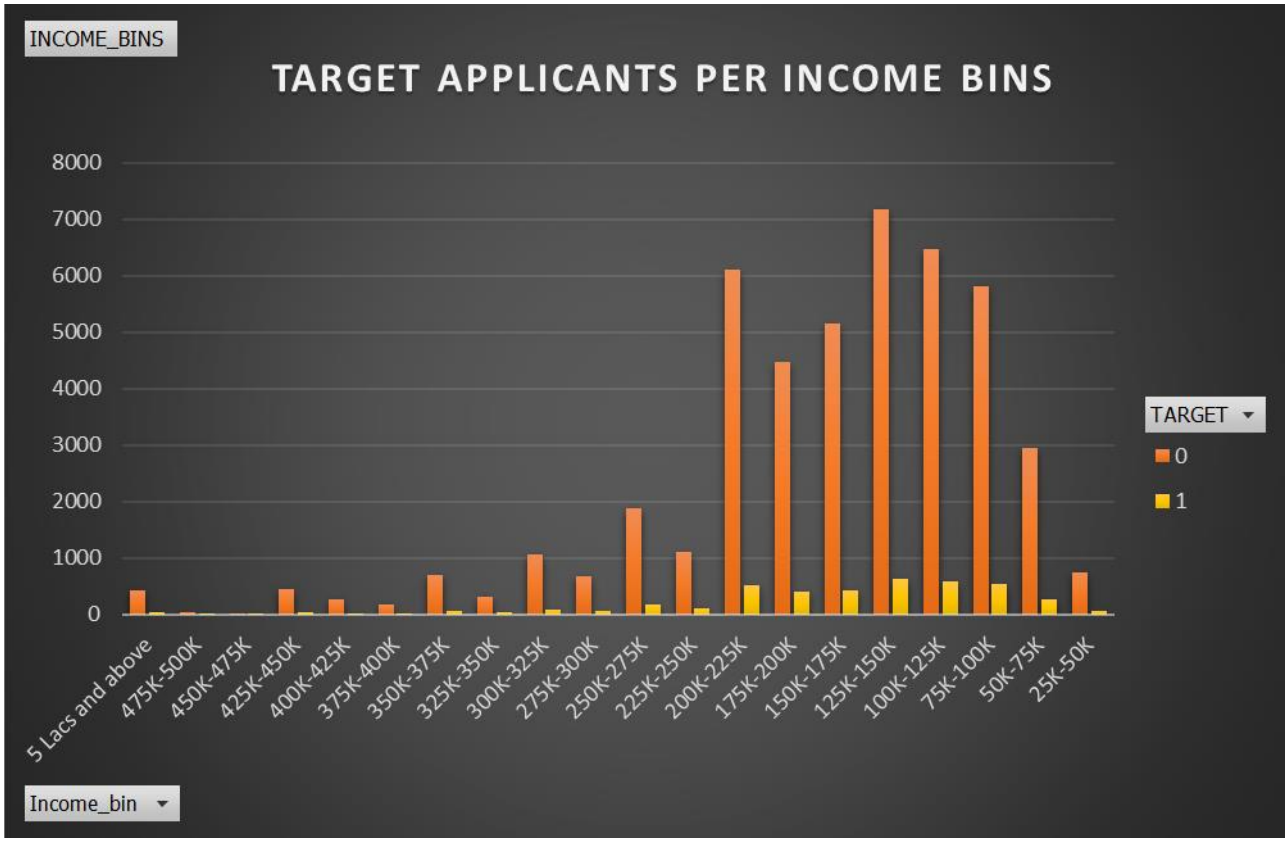
- The ratio of the number of instances with a target value of 0 to the number of instances with a target value of 1 is approximately 11.41.
- This ratio suggests a significant data imbalance, with approximately 11.41 times as many instances of category 0 as there are of category 1 in the dataset.

A ratio of 11.41 for the category 0 means that there are approximately 11.41 times as many instances of category 0 as there are of category 1 in your dataset. This suggests a significant imbalance in your data, with a much higher prevalence of category 0 compared to category 1. Imbalanced data can have implications for data analysis and modeling, so it's important to be aware of this when working with such data

Task 4: Univariate and Bivariate analysis

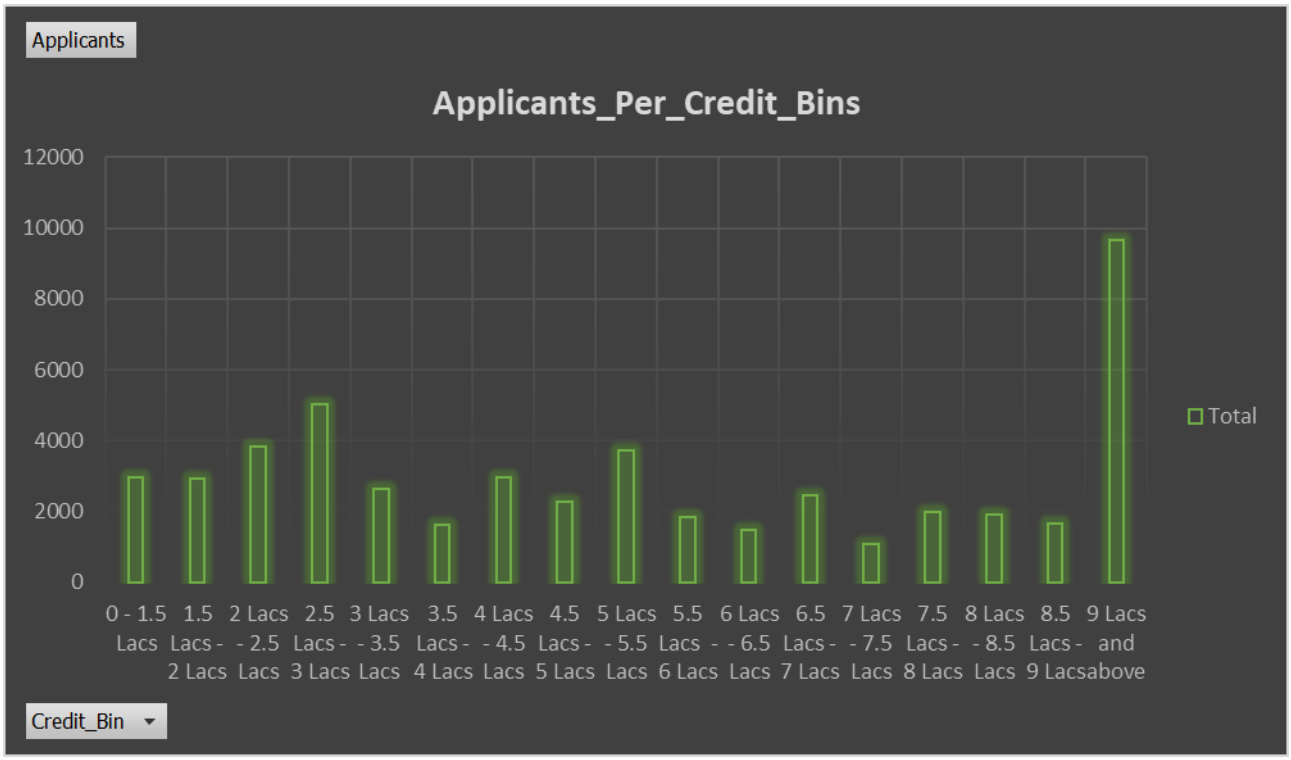
Univariate Analysis – TARGET APPLICANT PER INCOME_BINS:

- We created income bins to categorize applicants' income levels.
- The analysis revealed that the majority of applicants fall into the income range of "5 Lacs and above," with 454 applicants in this category.
- The "25K-50K" income range had the lowest number of applicants, with 804 applicants.
- This analysis provides a clear distribution of applicants based on income categories, which can help in understanding the income diversity among loan applicants.



Univariate Analysis – Applicants Per Credit Bins:

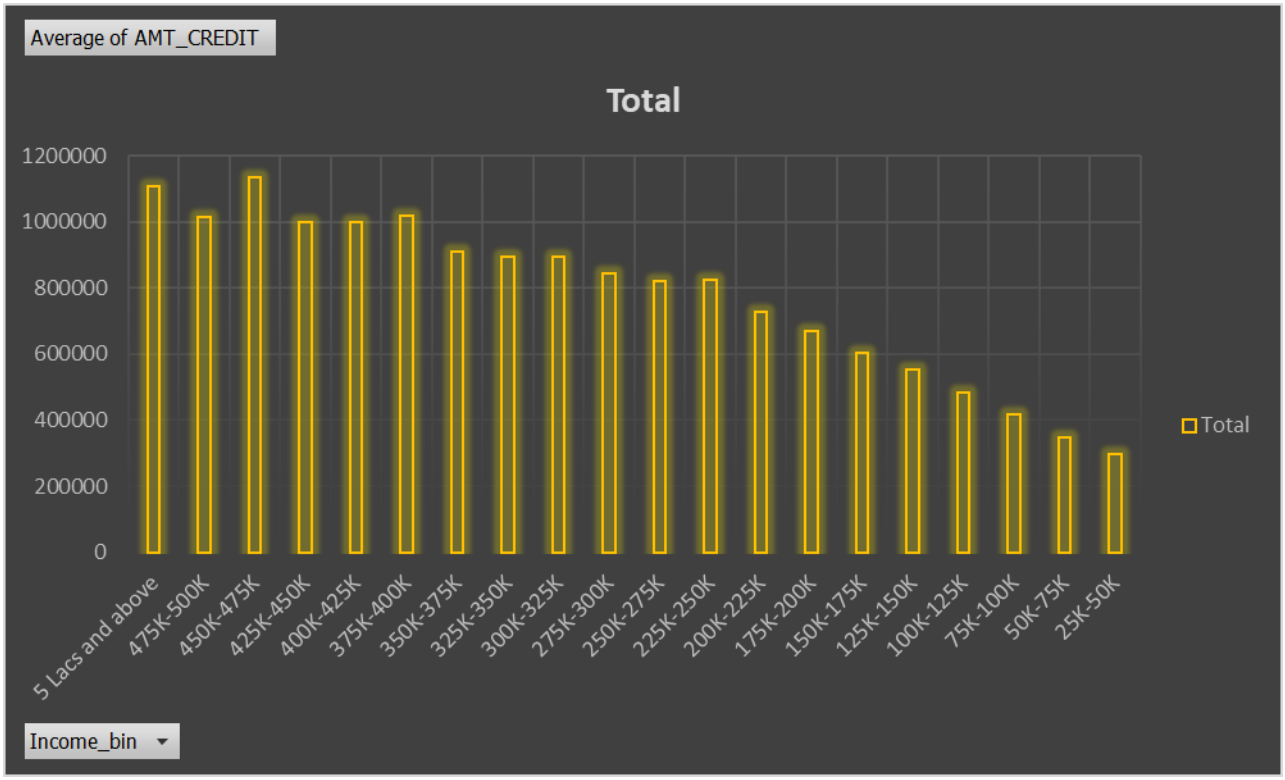
- We categorized applicants into income ranges, from "0 - 1.5 Lacs" to "9 Lacs and above."
- The highest number of applicants falls into the "9 Lacs and above" income range, with 9,670 applicants, while the "3 Lacs - 3.5 Lacs" income range has the lowest number of applicants, with 2,634 applicants.
- This analysis gives an overview of the distribution of applicants across different income brackets.



Bivariate Analysis - Average AMT_CREDIT by Income Range:

- We calculated the average AMT_CREDIT (credit amount) for each income range.
- Applicants in the "5 Lacs and above" income range have the highest average credit amount, with approximately 1,105,365.12.

- Applicants in the "25K-50K" income range have the lowest average credit amount, with approximately 297,752.08.
- This analysis shows how the average credit amount varies for different income categories, which can help understand the relationship between income and credit.



These analyses provide valuable insights into the distribution of loan applicants based on income, the number of applicants in different income ranges, and the average credit amount for each income range. Understanding these factors can be crucial for decision-making in loan processing and risk assessment.

1. Income Distribution and Customer Base:
    - The analysis reveals that the majority of loan applicants have incomes of "5 Lacs and above." This suggests that the lending institution's customer base primarily consists of individuals with relatively high incomes. Understanding this income distribution is crucial for tailoring loan products and services to the institution's target market.
2. Applicant Diversity:
    - The analysis highlights the diversity of applicants across different income ranges. This diversity is a positive sign for the institution, as it indicates a broad customer base. Serving clients from various income backgrounds can help reduce risk exposure and enhance the institution's resilience to economic fluctuations.
3. Credit Amount and Income:
    - Examining the average credit amounts in relation to income categories provides valuable insights. It shows that higher-income applicants tend to request larger loans. For the business, this information is essential for setting credit limits, designing loan products, and pricing strategies. It helps in aligning loan offerings with the financial capacity of applicants.
4. Risk Assessment and Income Levels:
    - Income distribution can directly impact risk assessment. Typically, higher-income applicants may be considered lower risk due to their financial stability. On the other hand, lower-income applicants may require more thorough scrutiny. Recognizing these variations allows the institution to implement risk management practices tailored to each income group.
5. Market Segmentation:

- Understanding the income distribution enables effective market segmentation. The institution can create targeted marketing campaigns and product offerings for specific income categories. This segmentation optimizes marketing efforts and enhances customer acquisition.

6. Profitability Analysis:

   - The analysis of credit amounts by income categories helps assess the institution's profitability. Higher-income applicants who seek larger loans can contribute significantly to interest income. However, it's important to balance profitability with risk to ensure sustainable growth.

7. Customized Loan Products:

   - Leveraging insights from income distribution, the institution can design customized loan products to cater to different income levels. Tailored products enhance customer satisfaction and increase the likelihood of loan approvals.

In summary, these insights enable the lending institution to make informed business decisions. They are crucial for risk management, product development, marketing strategies, and overall business planning. Understanding the income distribution and its correlation with credit amounts is fundamental to maintaining a balanced and profitable loan portfolio while meeting the diverse needs of customers.

Task 5: Identify Top Correlations for Different Scenarios:

Correlation 0:

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.036319722	0.005705458	0.02638396	0.001518097	-0.024912809	0.335876269	-0.243591518	-0.032537221	0.021288992
AMT_INCOME_TOTAL	0.036319722	1	0.377965752	0.451135167	0.384575912	0.181941261	0.073769425	-0.162702675	0.032286356	-0.205031899
AMT_CREDIT	0.005705458	0.377965752	1	0.770772818	0.986999774	0.095539444	-0.051084182	-0.077367219	-0.008290189	-0.102556478
AMT_ANNUITY	0.02638396	0.451135167	0.770772818	1	0.775835204	0.11727925	0.009911417	-0.113005288	0.00942697	-0.129920896
AMT_GOODS_PRICE	0.001518097	0.384575912	0.986999774	0.775835204	1	0.098968202	-0.048773297	-0.075106232	-0.00938552	-0.104841672
REGION_POPULATION_RELATIVE	-0.024912809	0.181941261	0.095539444	0.11727925	0.098968202	1	-0.030435419	-0.006610653	-0.002236288	-0.539333113
DAYS_BIRTH	0.335876269	0.073769425	-0.051084182	0.009911417	-0.048773297	-0.030435419	1	-0.615289978	0.270073313	0.00902485
DAYS_EMPLOYED	-0.243591518	-0.162702675	-0.077367219	-0.113005288	-0.075106232	-0.006610653	-0.615289978	1	-0.27222439	0.040505636
DAYS_ID_PUBLISH	-0.032537221	0.032286356	-0.008290189	0.00942697	-0.00938552	-0.002236288	0.270073313	-0.27222439	1	-0.008097427
REGION_RATING_CLIENT	0.021288992	-0.205031899	-0.102556478	-0.129920896	-0.104841672	-0.539333113	0.00902485	0.040505636	-0.008097427	1

Correlation 1:

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.010110177	0.007601905	0.029172977	-0.001079665	-0.020359154	0.2496732	-0.189324184	-0.042360717	0.055515557
AMT_INCOME_TOTAL	0.010110177	1	0.015271444	0.018004594	0.013269502	-0.006180303	0.009033662	-0.011555963	-0.009122006	-0.012846697
AMT_CREDIT	0.007601905	0.015271444	1	0.749665201	0.982267963	0.067775624	-0.142506035	0.016039571	-0.043771901	-0.045024534
AMT_ANNUITY	0.029172977	0.018004594	0.749665201	1	0.74950403	0.073123998	-0.008751713	-0.079556008	-0.02132109	-0.061578289
AMT_GOODS_PRICE	-0.001079665	0.013269502	0.982267963	0.74950403	1	0.076635488	-0.141005898	0.020235348	-0.049723232	-0.051296281
REGION_POPULATION_RELATIVE	-0.020359154	-0.006180303	0.067775624	0.073123998	0.076635488	1	-0.016468731	0.007742909	-0.005118563	-0.430032303
DAYS_BIRTH	0.2496732	0.009033662	-0.142506035	-0.008751713	-0.141005898	-0.016468731	1	-0.581479041	0.247896571	0.045027112
DAYS_EMPLOYED	-0.189324184	-0.011555963	0.016039571	-0.079556008	0.020235348	0.007742909	-0.581479041	1	-0.230063668	-0.009145883
DAYS_ID_PUBLISH	-0.042360717	-0.009122006	-0.043771901	-0.02132109	-0.049723232	-0.005118563	0.247896571	-0.230063668	1	0.025335227
REGION_RATING_CLIENT	0.055515557	-0.012846697	-0.045024534	-0.061578289	-0.051296281	-0.430032303	0.045027112	-0.009145883	0.025335227	1

1. Age and Loan Default Risk:

- For clients without payment difficulties, there's a small positive correlation between age and the number of children (0.036). This suggests that age is not a strong predictor of family size in this group.
- Understanding that age is not strongly linked to family size can influence lending decisions. It means that the age of applicants may not significantly impact their ability to repay loans, and the business can focus on more pertinent factors, such as income and credit history, when evaluating creditworthiness.

2. Income and Loan Amount:

- There is a moderate positive correlation (0.378) between income and the credit amount for clients without payment difficulties.
- This correlation is a vital piece of information for the lending institution. It means that clients with higher incomes tend to seek larger loans. This insight allows the business to set appropriate credit limits for different income brackets, tailoring loan products to suit the financial capacity of customers.

3. Credit Amount and Loan Annuity:

- A strong positive correlation (0.771) is observed between the credit amount and the annuity (loan installment) for these clients.
- This correlation highlights the natural financial relationship between loan amounts and associated annuity payments. Recognizing this relationship, the business can structure loans with installment plans that align with the loan amount, making it more manageable for clients.

4. Goods Price and Credit Amount:

- There's a very strong positive correlation (0.987) between the price of goods and the credit amount for clients without payment difficulties.
- This correlation signifies that clients typically request credit amounts closely matching the price of the goods they intend to purchase. For the business, this means that loan products can be designed to accommodate specific purchase needs, enhancing the overall customer experience.

5. Age and Employment Status:

- In this group, there's a strong negative correlation (-0.616) between age and the number of days employed.
- This insight indicates that younger clients tend to have shorter employment histories. The business can use this information to consider employment stability as a valuable factor in assessing credit risk. It may lead to offering more favorable terms to individuals with stable job histories.

6. Population Density and Region Rating:

- The negative correlation (-0.539) between population density and region rating suggests that regions with higher population densities tend to have lower ratings.
- This information can help the business in regional risk assessment and lending strategy. Regions with high population density may require specific risk management practices and loan product adjustments to account for higher competition for resources and potentially different economic dynamics.

7. Income and Age at ID Document Change:

- There's a slight positive correlation (0.032) between income and the age at which an ID document was last changed for clients without payment difficulties.
- While this correlation is relatively weak, it can still be useful for identity verification and fraud prevention. It's important to ensure that applicants with frequent changes in identity documents are thoroughly screened. This helps protect the business against identity-related risks.

In summary, these insights from the Excel analysis offer a range of advantages for the lending institution:

- Risk Assessment: Understanding the influence of age, income, and employment history on loan defaults helps in more accurate risk assessment. This, in turn, enables the business to make informed lending decisions.
- Product Customization: Insights about the relationship between income, loan amounts, and purchase prices enable the business to tailor loan products to better match customer needs.
- Regional Strategies: Recognizing the impact of population density on region ratings allows for the development of regional lending strategies and risk management practices.
- Identity Verification: Identifying the correlation between income and the age at which identity documents are updated can improve identity verification procedures, reducing the risk of fraud.

Overall, these insights empower the business to make data-driven decisions, manage risk effectively, and offer tailored financial products to its customers, enhancing the overall performance and profitability of the lending institution.

5. Result:

Through this project, a comprehensive analysis of the Bank Loan Case Study dataset has been conducted, leading to valuable insights and a deeper understanding of various aspects related to loan applications and customer characteristics. The project's achievements can be summarized as follows:

1) Data Preprocessing: The project successfully prepared and cleaned the raw dataset, addressing missing values and inconsistencies. This step was crucial to ensure the quality and reliability of the subsequent analysis.
2) Outlier Identification: Outliers in the dataset were detected using Excel's statistical functions, allowing for a better understanding of extreme data points. These outliers were assessed for their validity and impact on the analysis.
3) Data Imbalance Analysis: The project assessed data imbalance within the loan application dataset. It was found that there is a significant imbalance, with a much higher prevalence of category 0 (no payment difficulties) compared to category 1 (payment difficulties). This understanding is crucial for building reliable models and assessing the impact of imbalanced data on analysis outcomes.
4) Univariate and Bivariate Analysis: Univariate, segmented univariate, and bivariate analyses were performed to gain insights into the driving factors of loan default. These analyses provided a deep understanding of individual variables, relationships between variables, and their associations with loan default.
5) Correlation Analysis: The project conducted correlation analyses for different scenarios, revealing the top correlations for each segmented dataset. This helped identify strong indicators of loan default, such as age, income, credit amount, and regional factors.
6) Business Insights: The analysis results were translated into actionable business insights. For instance, it was observed that younger applicants are more likely to default on loans, suggesting the need for stricter approval criteria for this demographic. Additionally, a strong correlation between credit amount and the price of goods highlighted the importance of tailoring loan products based on the cost of goods customers intend to buy.
7) Risk Assessment: The insights from this project contribute significantly to risk assessment and lending policies. For example, the correlation between employment history and loan default indicates the importance of considering employment stability when evaluating credit risk.
8) Regional Considerations: The correlation between regional factors and loan default provides valuable information for regional risk assessment and the development of region-specific lending strategies.

Overall, this project has enhanced our understanding of the Bank Loan Case Study and provided actionable insights that can inform decision-making in various fields, including risk management, credit assessment, and product development. These findings contribute to more informed, data-driven strategies for the bank's loan operations and can help mitigate risk while providing better services to customers.