CREDIT EDA ASSIGNMENT

BY -
AASTHA SHARMA

PROBLEM STATEMENT

1. Companies face significant challenges in loan approval due to applicants' insufficient or non-existent credit histories, which can lead to increased risks of default. This case study focuses on using Exploratory Data Analysis (EDA) to scrutinize patterns within loan application data to ensure capable applicants are not unjustly rejected. The data encompasses various scenarios including clients with payment difficulties and those who maintain timely payments.

2. When assessing loan applications, the company faces two major risks: losing potential business from applicants likely to repay if loans are not approved, and incurring financial losses from approving loans for likely defaulters. The dataset categorizes applications into four decisions: Approved, Cancelled, Refused, and Unused offers, reflecting different outcomes based on both client decisions and company assessments. By analyzing consumer and loan attributes through EDA, the company aims to discern factors influencing default tendencies, thereby refining their approval processes to balance risk and opportunity effectively.

DATASETS OF INTEREST

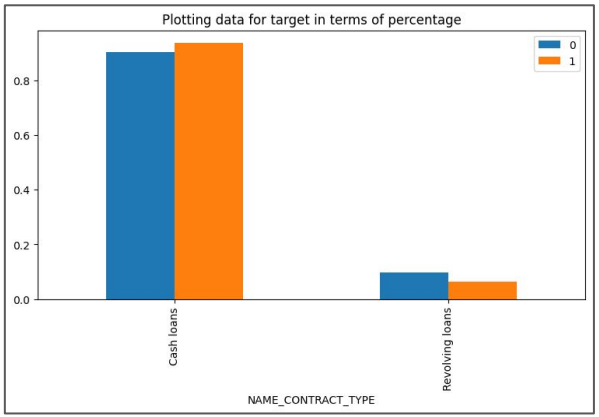
1. 'application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

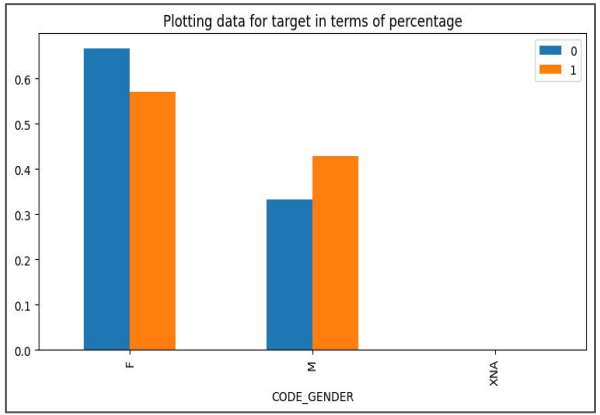
2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.

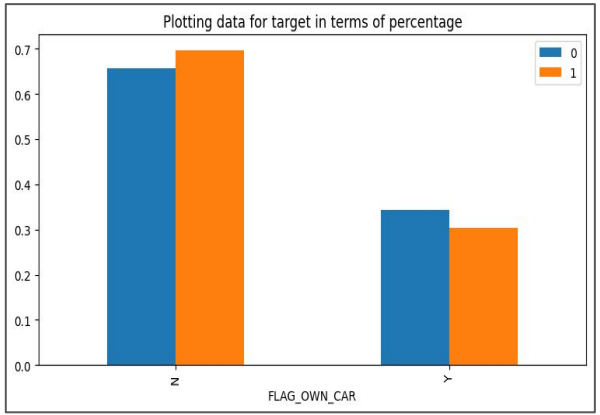
METHODOLOGY

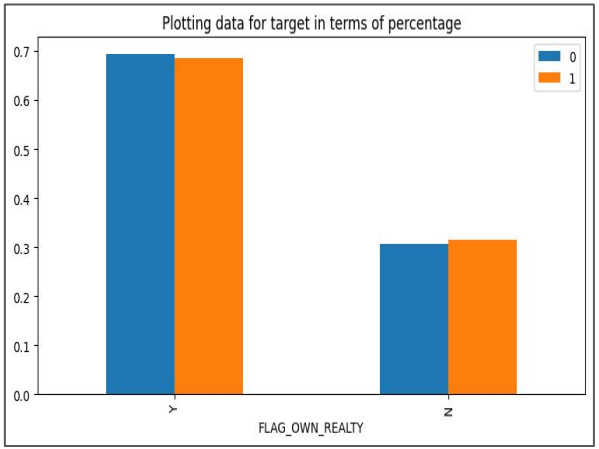
- Data Collection
- Loading and Studying the Data
- Data Cleaning:
  - Removing and Imputing null values
  - Removing unnecessary columns
- Data Standardisation
  - Changing negative values to positive
  - Binning of columns
- Univariate Analysis:
  - Analysis of Numerical columns
  - Analysis of Categorical columna
- Bivariate Analysis:
  - Numerical - Categorical Analysis
  - Numerical - Numerical Analysis
- Multivariate Analysis:
  - HeatMap of Merged Datasets

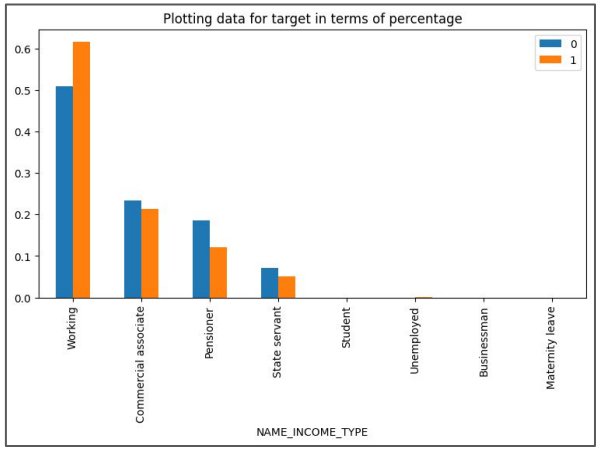
Univariate Analysis of *"application_data"*

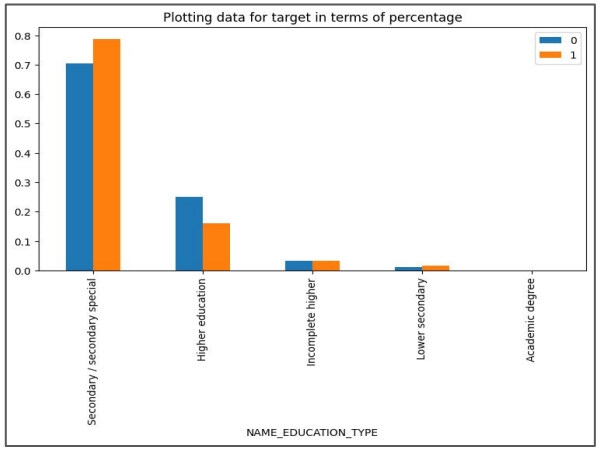








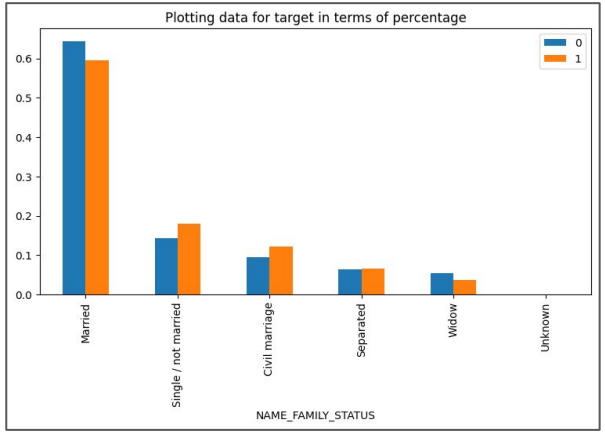


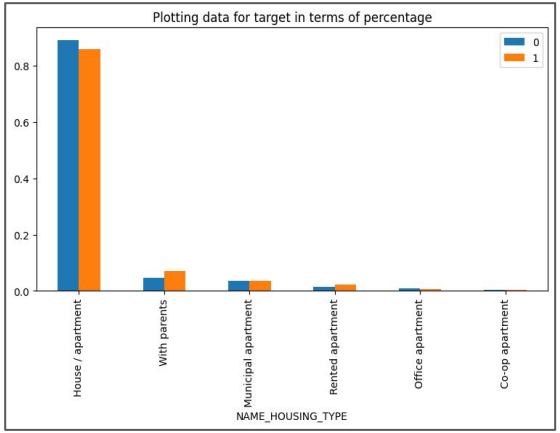


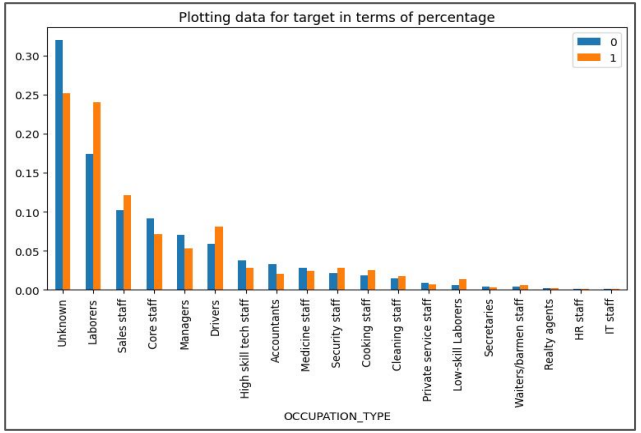
INSIGHTS

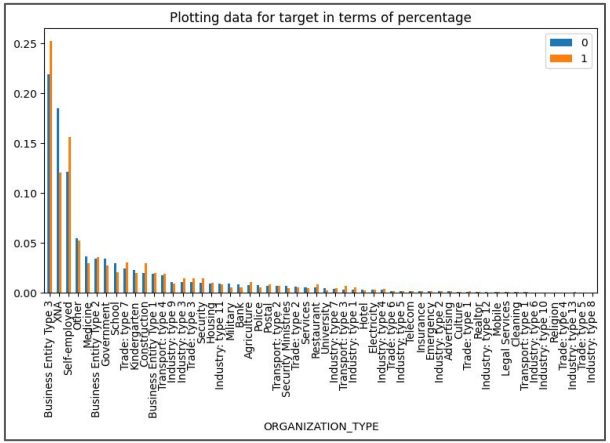
- Cash loans have a higher proportion of both target 0 and target 1 compared to revolving loans.
- Females have a higher proportion of both target 0 and target 1 compared to males.
- Individuals without a car have a higher proportion of both target 0 and target 1 compared to those with a car.
- Individuals who own real estate have a higher proportion of both target 0 and target 1 compared to those who do not own real estate.
- Working individuals have the highest proportion of both target 0 and target 1, with commercial associates and pensioners following.
- Individuals with secondary/special secondary education have the highest proportion of both target 0 and target 1, followed by those with higher education.

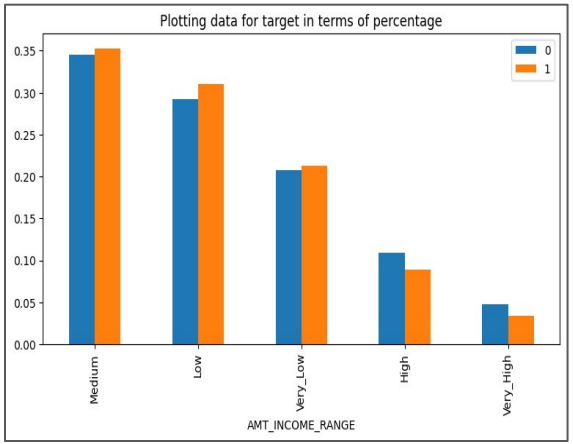
Univariate Analysis of *"application_data"*

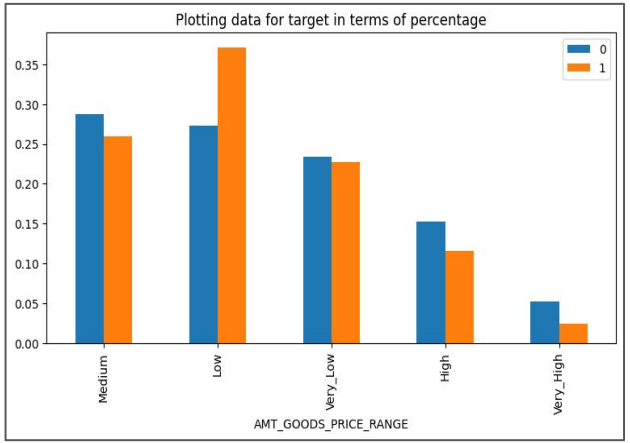








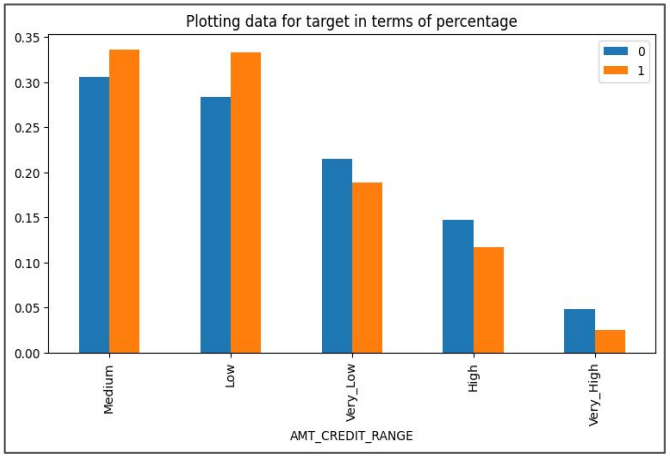


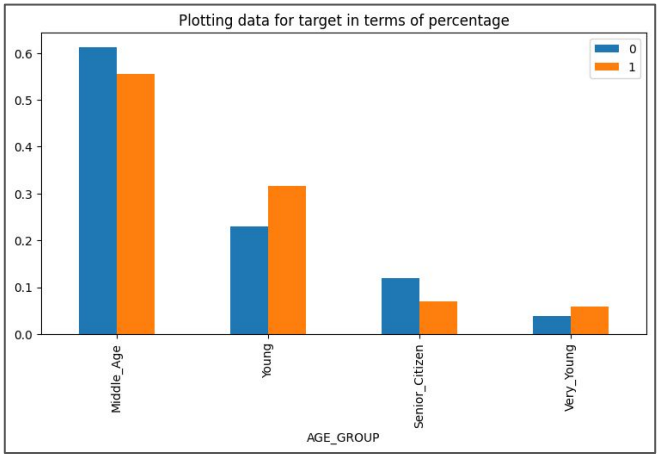


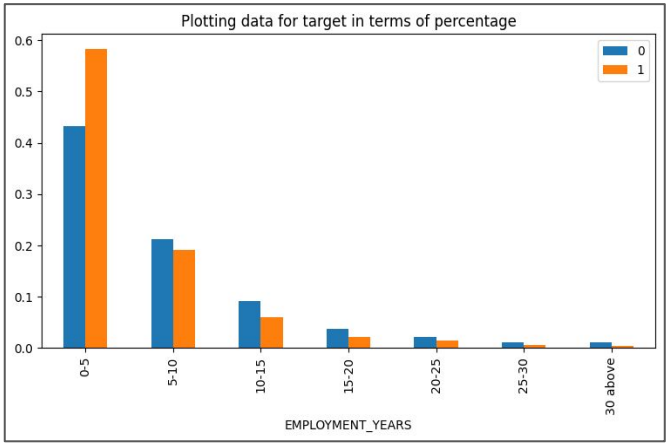
INSIGHTS

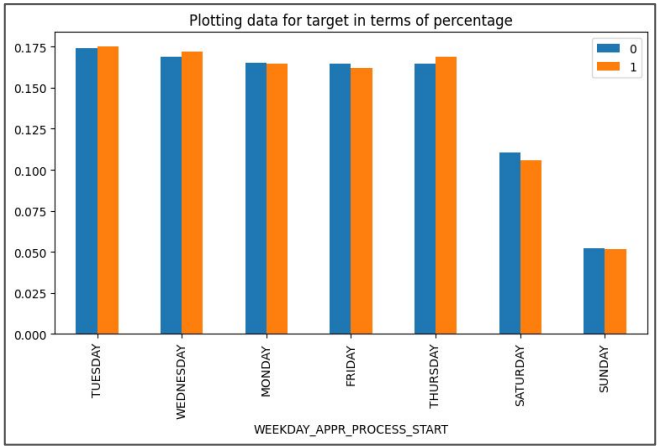
- Married individuals have the highest proportion of both target 0 and target 1, with single individuals following.
- House/apartment dwellers have the highest proportion of both target 0 and target 1, with negligible representation from other housing types.
- The 'Laborers' category has the highest proportion of both target 0 and target 1, followed by 'Sales staff' and 'Core staff'.
- Business entities have the highest proportion of both target 0 and target 1, with other categories showing minimal representation.
- Medium income range has the highest proportion of both target 0 and target 1, followed by low and very low income ranges.
- Medium goods price range has the highest proportion of both target 0 and target 1, with low and very low price ranges following.

Univariate Analysis of *"application_data"*





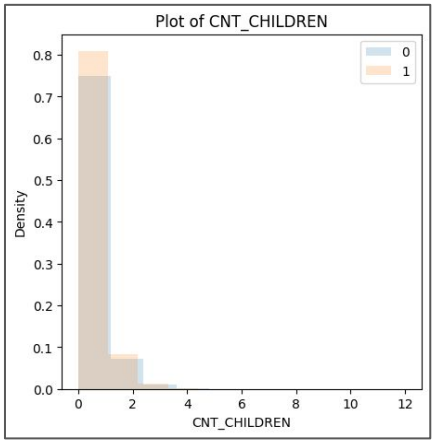


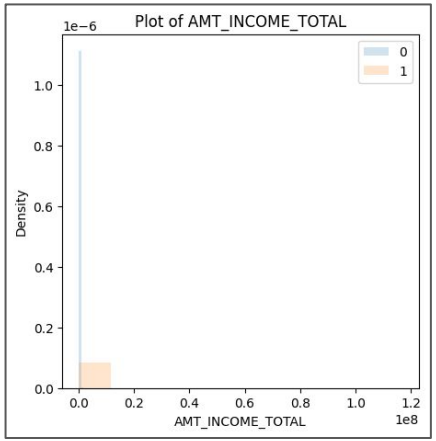


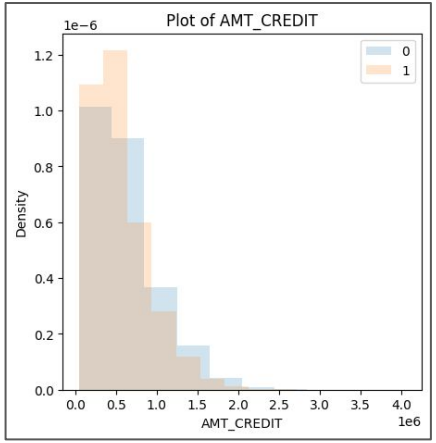
INSIGHTS

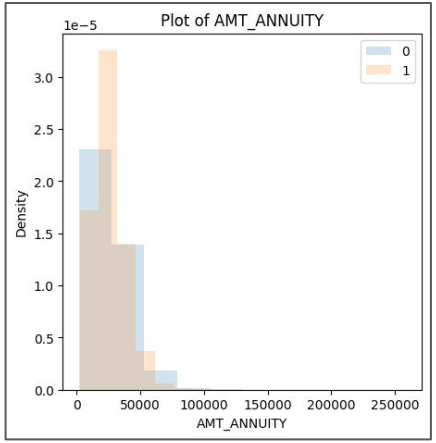
- Distribution of the target variable shows that the vast majority of applicants (around 95%) do not have payment difficulties.
- Distribution of loan amount shows that there is a wide range of loan amounts, with some applicants applying for very small loans and others applying for very large loans.
- Distribution of applicant age suggests that younger applicants are more likely to apply for loans than older applicants.
- Distribution of employment years suggests that applicants who have been employed for a longer period of time are more likely to apply for loans than those who have been employed for a shorter period of time.
- Distribution of day of the week that the application process started suggests that there may be more loan applications started on weekdays than on weekends.

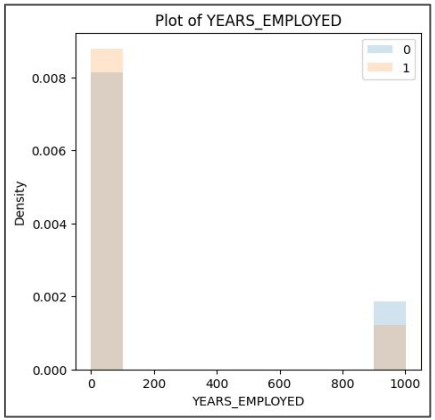
Univariate Analysis of *"application_data"*

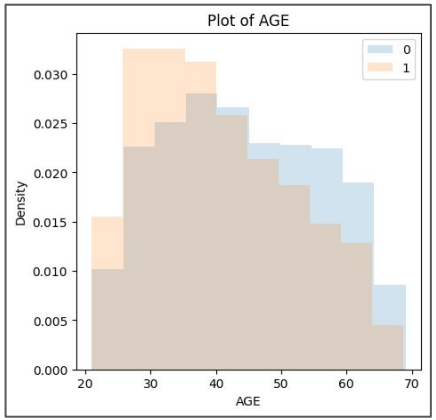


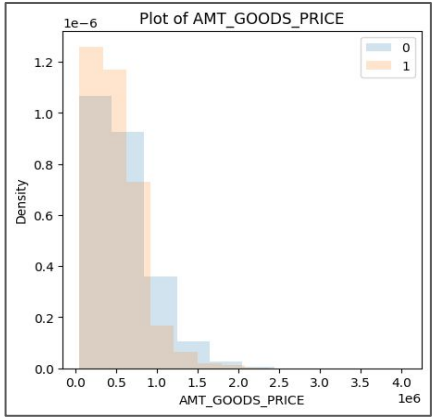


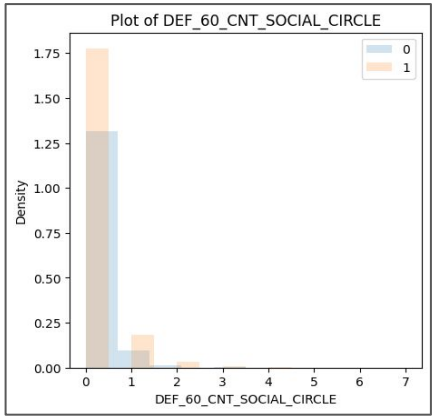








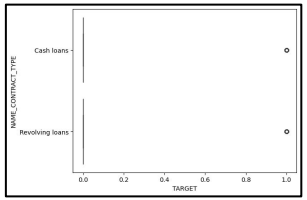


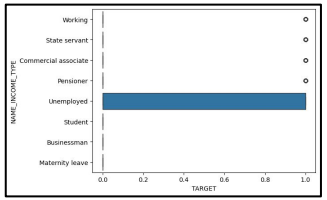


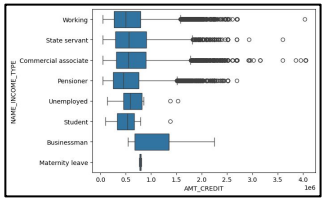
INSIGHTS

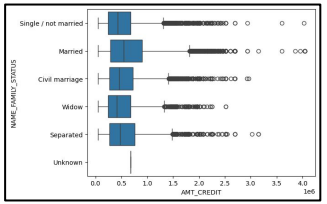
- Number of children distribution is right skewed, with most applicants having zero children.
- Income distribution is right skewed, with most applicants having a lower income.
- Loan amount distribution is right skewed, with most applicants having a lower loan amount.
- Credit amount distribution is right skewed, with most applicants having a lower credit amount.
- Years employed distribution is right skewed, with most applicants having fewer years of employment.
- Age distribution appears to be unimodal, with a peak around 25-30 years old.
- Goods price distribution is right skewed, with most applicants having a lower goods price.
- Social circle distribution is right skewed, with most applicants having a smaller social circle.

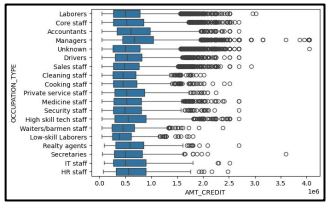
Bivariate Analysis of *“application_data”*

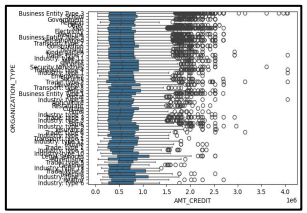


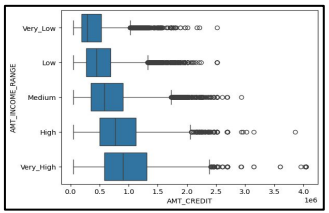


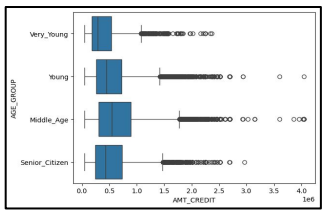


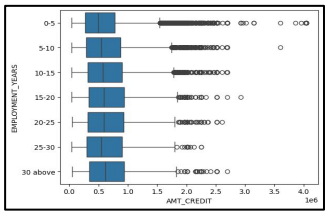


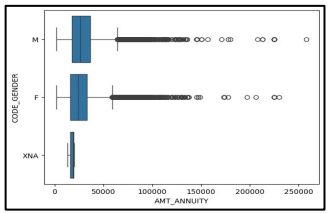


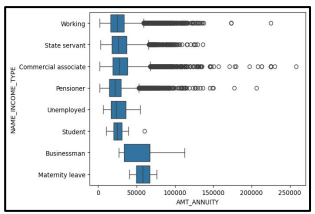


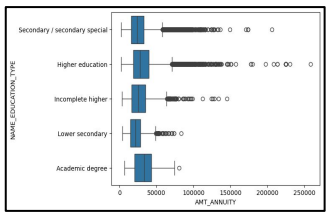


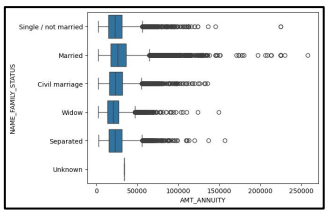


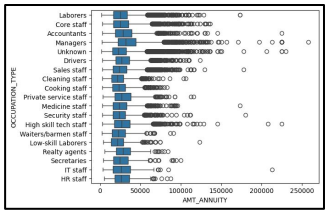


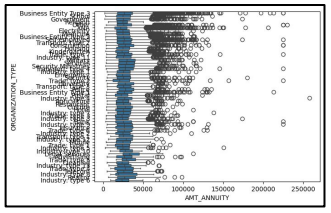


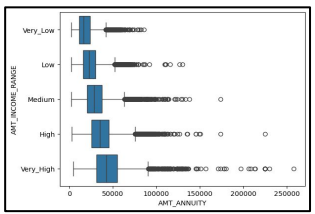


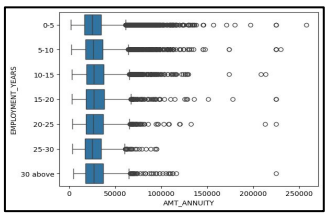


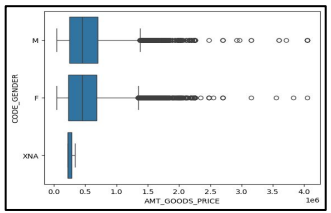


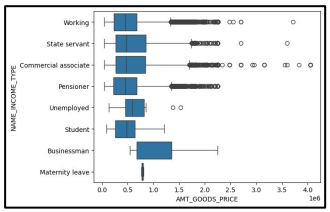


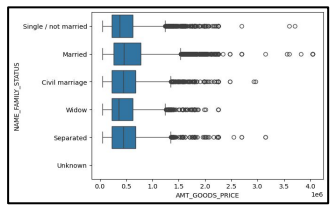








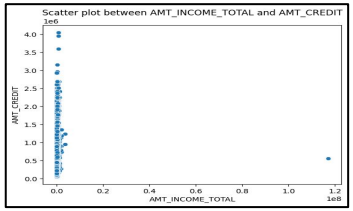


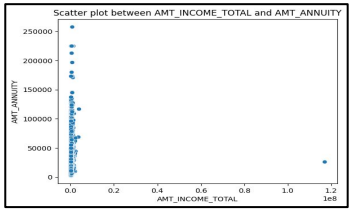


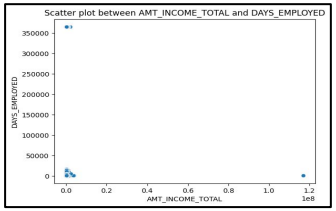
INSIGHTS

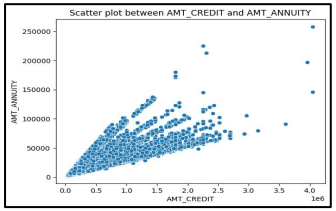
- Cash loans show a clear distinction in target variable compared to revolving loans.
- Unemployed individuals have a higher target value.
- Most single/not married individuals have a higher target value.
- Business entities and laborers have higher credit amounts.
- Higher education correlates with higher credit amounts.
- Middle-aged individuals tend to have higher credit amounts.
- Higher number of children correlates with higher target value.
- Longer employment duration shows a lower target value.
- Males have higher goods prices than females.
- Males have higher annuity amounts than females.
- Married individuals have higher annuity amounts.
- Business entities have higher goods prices.
- Higher education correlates with higher goods prices.
- Older individuals tend to have higher goods prices.
- Males have higher goods prices than females. Business entities and laborers have higher goods prices.
- Married individuals tend to have higher goods prices.
- Higher education correlates with higher goods prices.
- Older individuals tend to have higher goods prices.

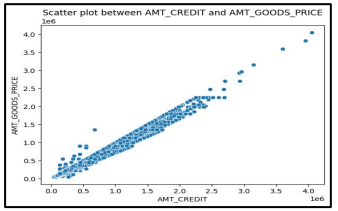
Bivariate Analysis of *"application_data"*

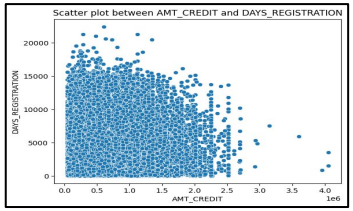


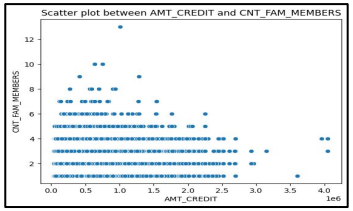


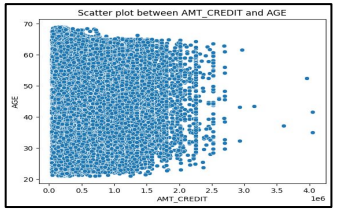


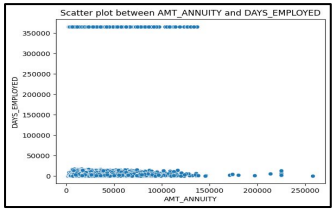


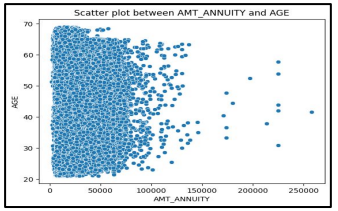


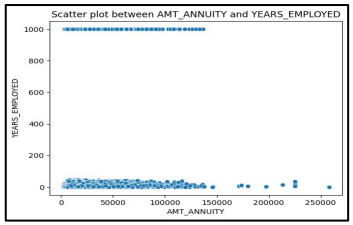


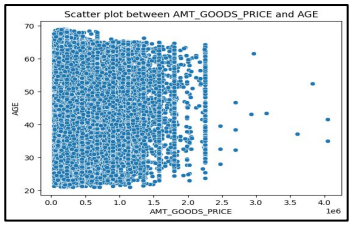


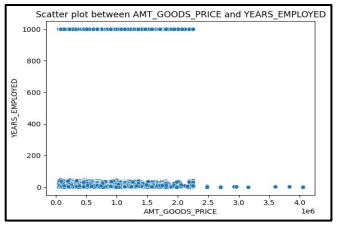


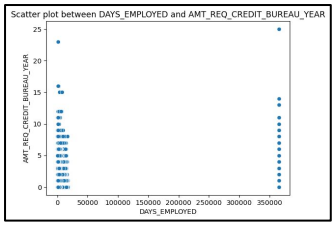


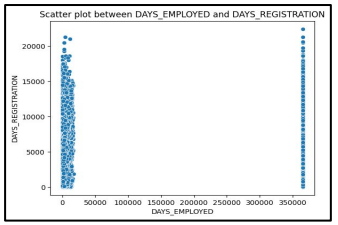








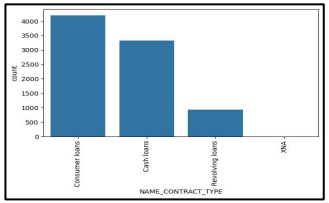


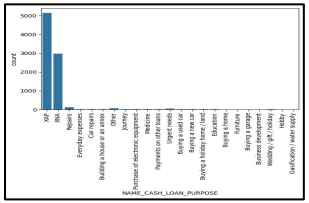


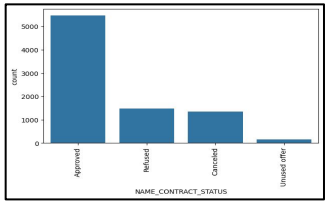
INSIGHTS

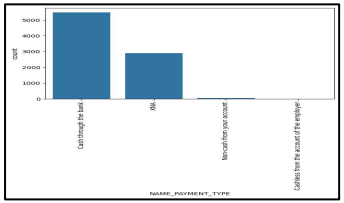
- Credit amount increases with total income but with a wide spread.
- Higher total income slightly correlates with higher annuity amounts.
- Most data points cluster at lower income and days employed values.
- Annuity amount increases with credit amount.
- Strong positive correlation between credit amount and goods price.
- Days since registration shows no clear pattern with credit amount.
- Family members count has no clear relationship with credit amount.
- Credit amount is uniformly distributed across different ages.
- Higher annuity amounts correspond to more days employed but with a wide spread.
- Annuity amount is uniformly distributed across different ages.
- Years employed shows no clear pattern with annuity amount.
- Goods price is uniformly distributed across different ages.
- No clear relationship between goods price and years employed.
- Most requests to credit bureau happen at lower days employed values.
- Longer days employed generally correspond to longer days since registration.

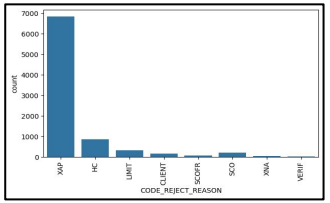
Univariate Analysis of *“previous_application”*

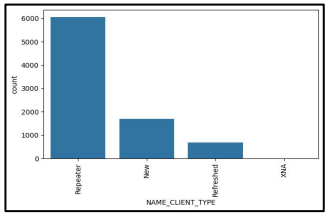


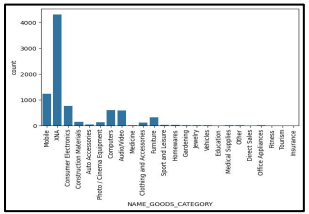


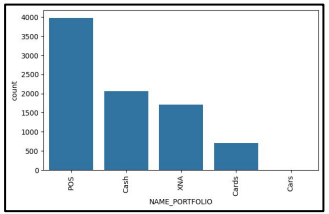


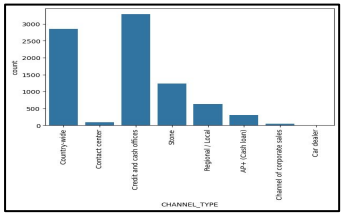


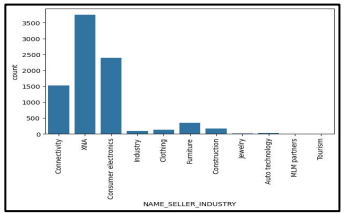








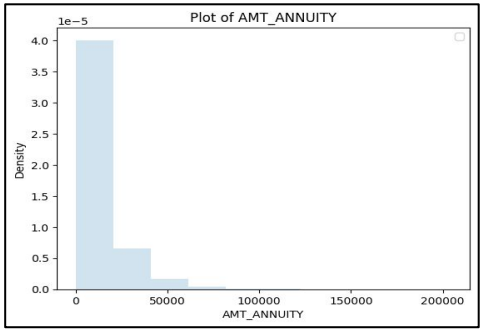


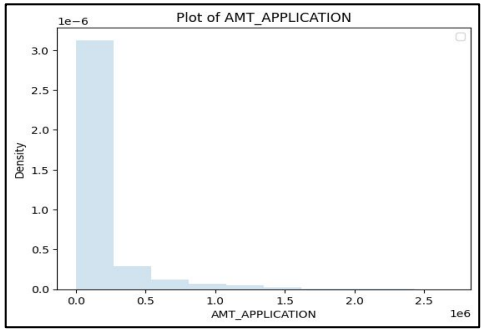


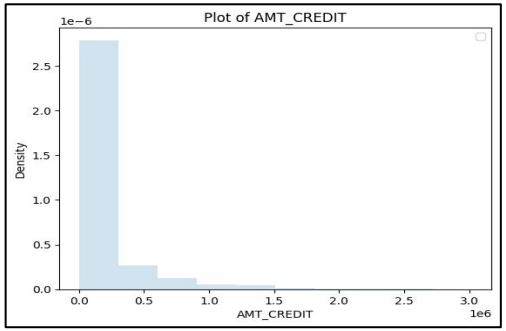
INSIGHTS

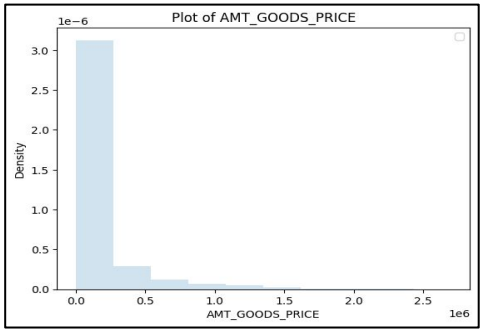
- Consumer loans are the most frequent, followed by cash loans. Revolving loans are the least common.
- "Repairs" are the top purposes for cash loans. Other purposes like "Building a house" and "Buy furniture" are less common.
- Approved contracts dominate the dataset. Refused contracts are also common, while canceled and unused offers are fewer.
- Cash payments through the bank are the most prevalent, other types like non-cash from own account have lower counts.
- Repeater clients form the majority, with new clients also notable. Refreshed clients are the least frequent.
- Mobile and Computers are the leading goods categories. Other categories like Audio/Video and Furniture have lower frequencies.
- POS portfolios are the most common, followed by Cash. Cards are less frequent.
- Country-wide channels are the most used, with Credit and Wealth offices also significant. Other channels like Stone and Regional/Warehouse are less common.
- Connectivity is the most common. Other industries like Consumer electronics, Industry, and Jewelry have fewer counts.

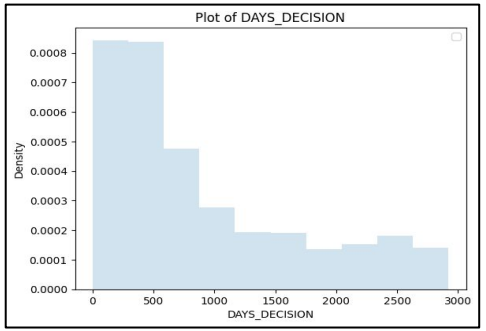
Univariate Analysis of *"previous_application"*

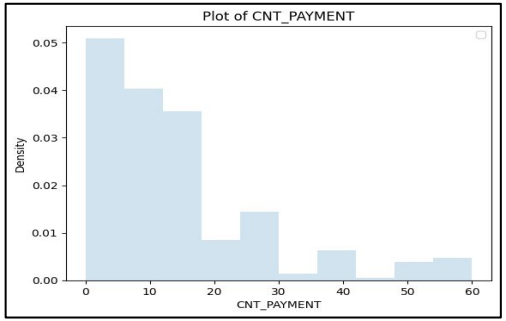








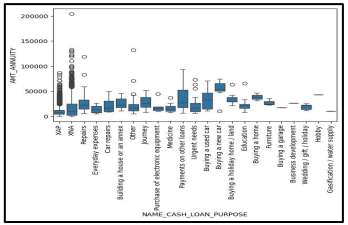


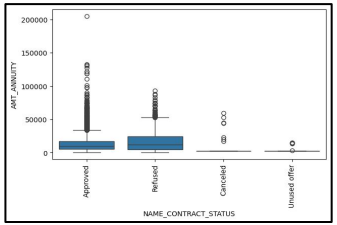


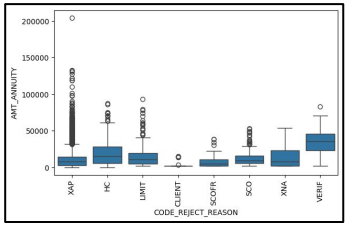
INSIGHTS

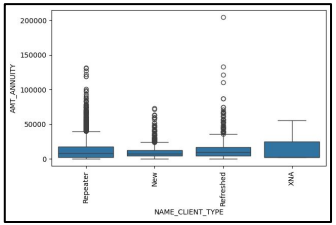
- The distribution is highly skewed to the right, with most annuities falling below 50,000. Very few annuities exceed 100,000.
- The majority of applications are for amounts below 500,000, showing a right-skewed distribution. Applications above 1,000,000 are rare.
- Most credit amounts are concentrated below 500,000, indicating a right-skewed distribution. Amounts above 1,000,000 are uncommon.
- Goods prices are predominantly below 500,000, with a right-skewed distribution. Prices exceeding 1,000,000 are infrequent.
- Decisions are mostly made within 1,000 days, with a decreasing frequency over time. Very few decisions are made after 2,500 days.
- The number of payments is largely concentrated between 0 and 20, showing a right-skewed pattern. Higher payment counts, such as 50 or 60, are rare.

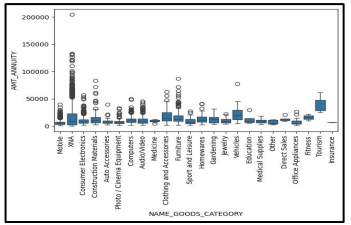
Bivariate Analysis of *“previous_application”*

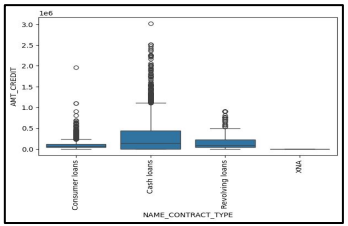


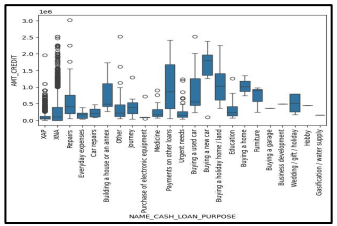


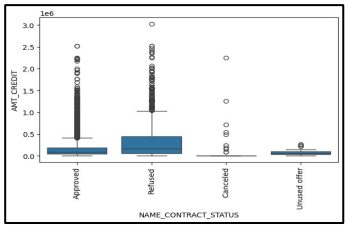


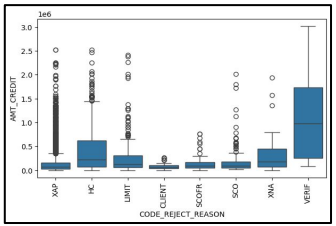


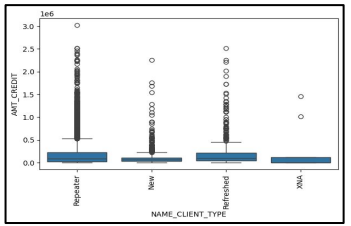








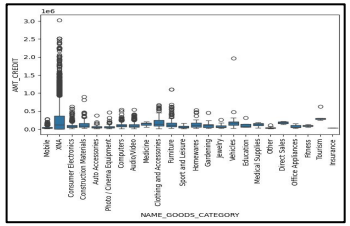


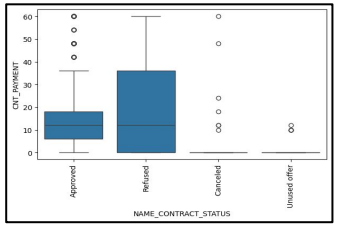


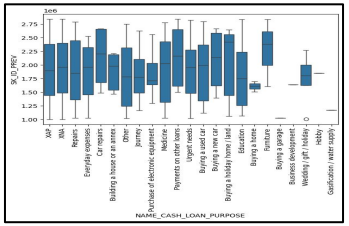
INSIGHTS

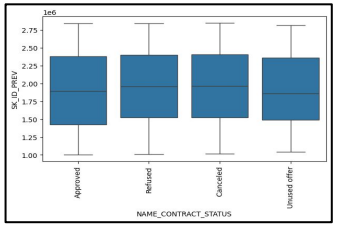
- Higher annuities are associated with purposes like "Repairs" and "Building a house". Categories like "Hobby" and "Journey" have lower, more consistent annuity values.
- Approved contracts show a wider range and higher median annuity values. Refused and canceled contracts typically have lower annuities.
- "XAP" (approved) has the highest range and median annuity. Other rejection reasons like "HC" and "LIMIT" show lower annuity values.
- Repeater clients exhibit higher and more variable annuity values. New clients have lower median annuities compared to repeaters.
- Categories like "Cars" and "Computers" have higher and more variable annuities. Categories such as "Clothing" and "Consumer Electronics" have lower annuities.
- "Cash loans" show a higher median and wider spread in credit amounts. "Revolving loans" generally have lower credit values.
- Higher credits are seen in purposes like "Repairs" and "Building a house". Lower credits are observed for purposes like "Hobby" and "Journey".
- Approved contracts have the highest range and median credit amounts. Refused and canceled contracts show lower credit amounts.
- "XAP" (approved) category has higher credit values. Other reasons for rejection generally correspond to lower credit values.
- Repeater clients have higher median and more variable credit amounts. New clients generally have lower median credit values.

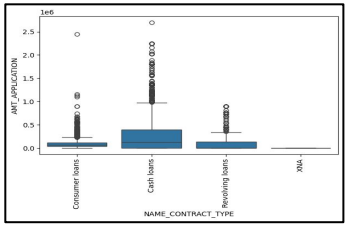
Bivariate Analysis of *“previous_application”*

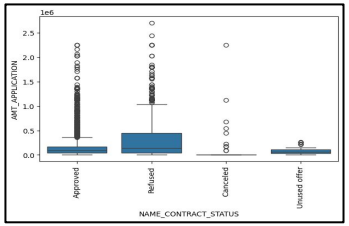


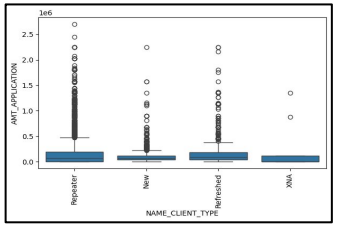


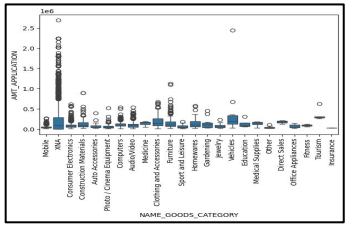


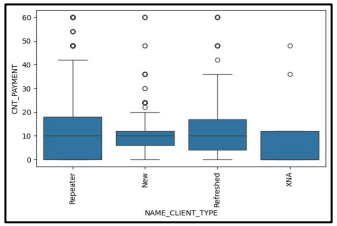


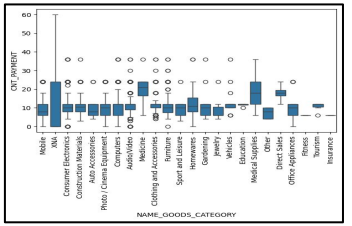








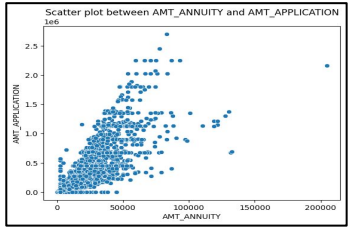


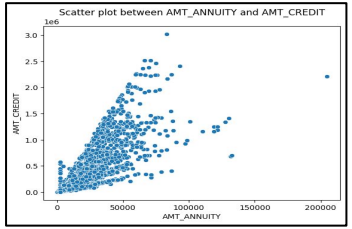


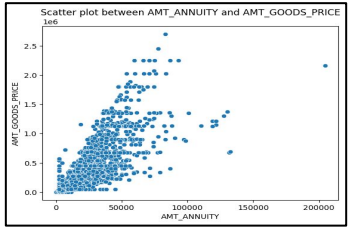
INSIGHTS

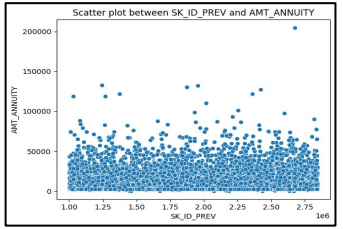
- Goods categories like "Mobile", "Computers", and "Cars" have higher credit amounts. Other categories such as "Clothing" and "Sporting goods" show lower credit values.
- Refused contracts have a higher count of payments compared to approved ones. Canceled contracts show a lower count of payments.
- Categories like "Repairs" and "Building a house" have higher goods prices. Lower goods prices are seen in purposes like "Hobby" and "Journey".
- Approved contracts have higher goods prices compared to other statuses. Refused and canceled contracts have lower goods prices.
- "Cash loans" have higher application amounts compared to "Consumer loans"."Revolving loans" generally have lower application amounts.
- Approved contracts show higher application amounts than refused ones. Canceled contracts have lower application amounts.
- Repeater clients exhibit higher application amounts. New clients have lower application amounts.
- Categories like "Mobile", "Computers", and "Cars" have higher goods prices. Lower goods prices are observed in categories like "Clothing" and "Sporting goods".
- Repeater clients have a higher count of payments compared to new clients. New clients show lower counts of payments.
- Categories like "Computers" and "Cars" show a higher count of payments. Lower counts of payments are seen in categories like "Clothing" and "Consumer Electronics".

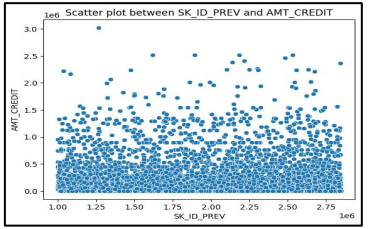
Bivariate Analysis of *“previous_application”*

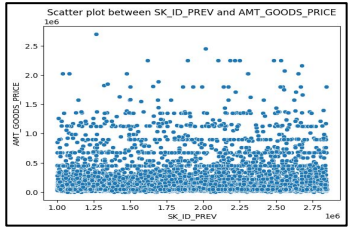


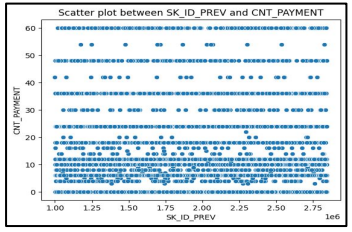


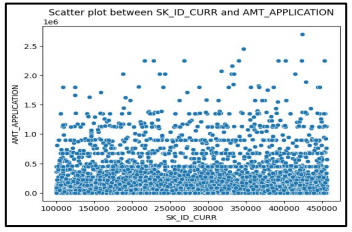


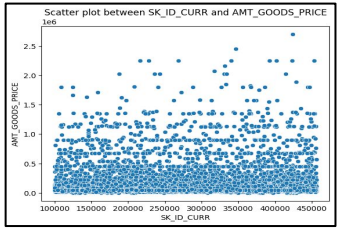


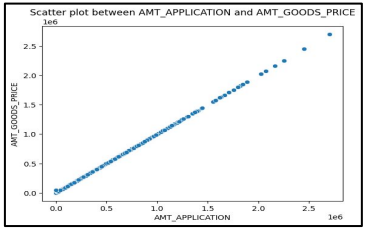


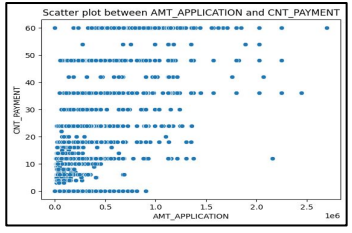


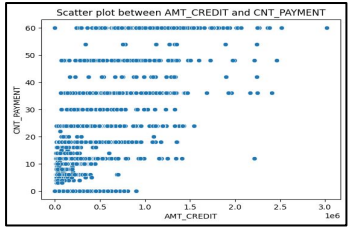


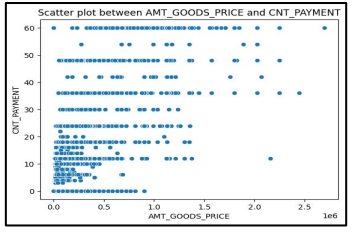


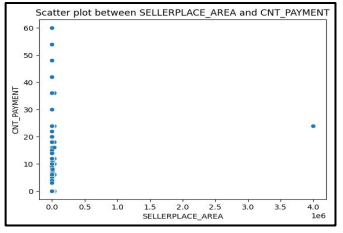


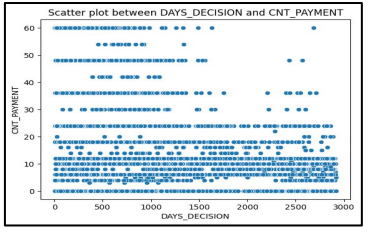












INSIGHTS

- There is a clear positive correlation between AMT_ANNUITY and AMT_APPLICATION, indicating that higher application amounts tend to be associated with higher annuity amounts. The data points form a triangular shape, suggesting a wide range of annuity amounts for lower application amounts, but more constrained values for higher application amounts.

- A positive correlation is observed between AMT_ANNUITY and AMT_CREDIT, implying that higher credit amounts tend to accompany higher annuity amounts. However, the spread of data points increases with higher credit amounts, indicating variability.

- The scatter plot shows a positive correlation between AMT_ANNUITY and AMT_GOODS_PRICE, suggesting that higher goods prices are associated with higher annuity amounts. The data points exhibit a similar triangular distribution as the first plot.

- There is no obvious correlation between SK_ID_PREV and AMT_ANNUITY, as the data points are widely scattered across the plot. This indicates that the previous application ID does not have a direct relationship with the annuity amounts.

- No clear correlation is found between SK_ID_PREV and AMT_CREDIT, with data points scattered uniformly. This suggests that the previous application ID does not influence the credit amounts.

- The scatter plot does not reveal any significant correlation between SK_ID_PREV and AMT_GOODS_PRICE, with data points evenly distributed. The previous application ID appears unrelated to the goods prices.

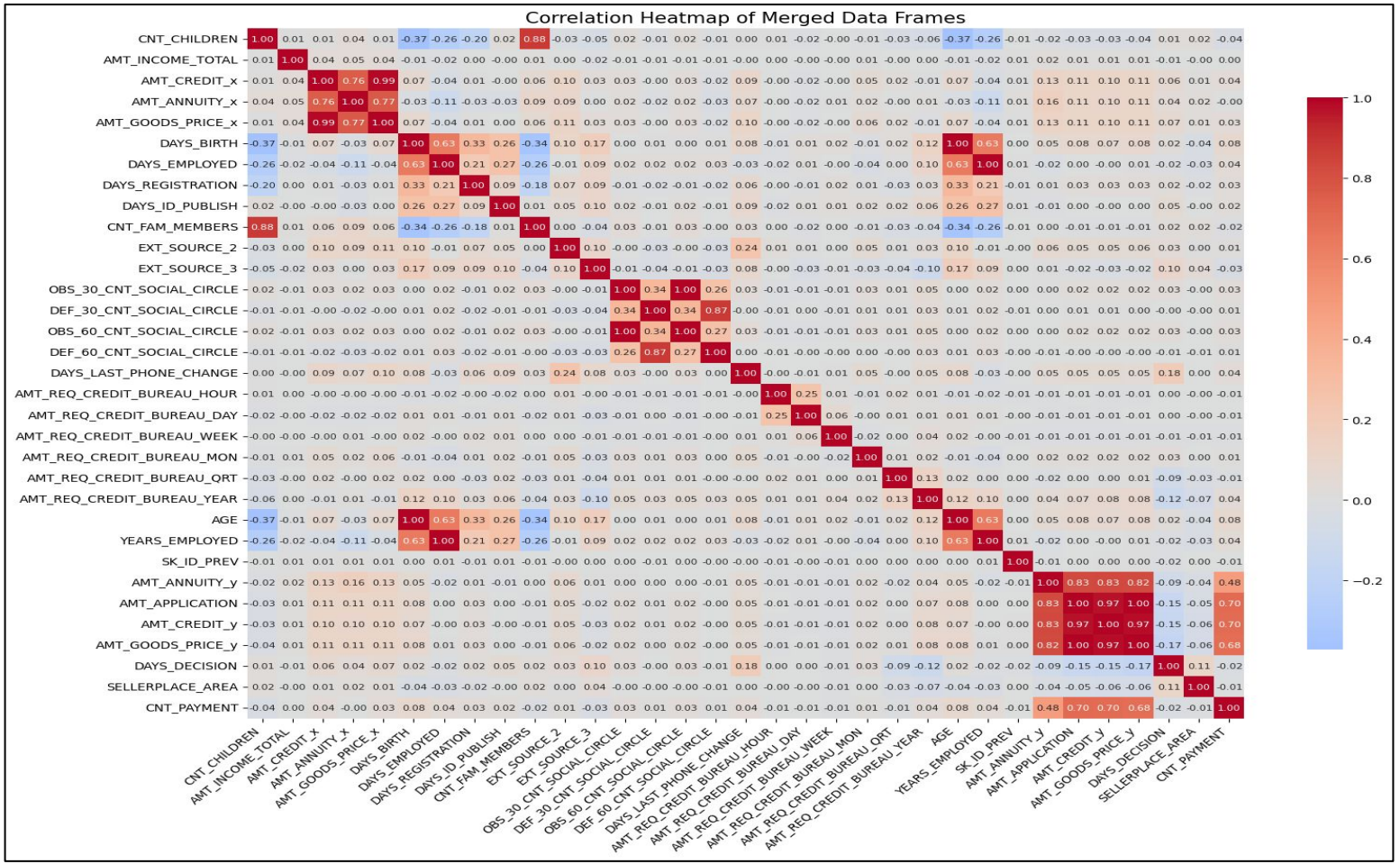
- There is no discernible correlation between SK_ID_PREV and CNT_PAYMENT, as data points are uniformly spread across the plot. This indicates that the previous application ID does not affect the number of payments.

- The plot shows no clear correlation between SK_ID_CURR and AMT_APPLICATION, with data points dispersed without a noticeable pattern. The current application ID does not seem to influence the application amounts.

INSIGHTS

- There is no evident correlation between SK_ID_CURR and AMT_GOODS_PRICE, with data points scattered randomly. This suggests that the current application ID does not impact the goods prices.
- A strong positive correlation is observed between AMT_APPLICATION and AMT_GOODS_PRICE, indicating that higher application amounts are consistently associated with higher goods prices. The relationship is almost linear.
- The scatter plot shows no clear correlation between AMT_APPLICATION and CNT_PAYMENT, with data points widely scattered. The number of payments does not appear to be influenced by the application amount.
- There is no significant correlation between AMT_CREDIT and CNT_PAYMENT, as data points are uniformly spread. This indicates that the number of payments is not related to the credit amount.
- No clear correlation is found between AMT_GOODS_PRICE and CNT_PAYMENT, with data points scattered across the plot. The number of payments does not seem to be influenced by the goods prices.
- The plot reveals no apparent correlation between SELLERPLACE_AREA and CNT_PAYMENT, with data points densely clustered at the lower values of SELLERPLACE_AREA. This suggests a lack of relationship between the seller place area and the number of payments.
- There is no obvious correlation between DAYS_DECISION and CNT_PAYMENT, as data points are uniformly distributed across the plot. The number of days taken for a decision does not seem to affect the number of payments.

Multivariate Analysis of Merged Data Frames



INSIGHTS

- AMT_APPLICATION, AMT_CREDIT, and AMT_GOODS_PRICE show high positive correlations with each other (around 0.99), indicating these variables are closely related and likely measure similar aspects of the loan application.
- EXT_SOURCE_2 and EXT_SOURCE_3 have a moderate negative correlation with AMT_ANNUITY, indicating that higher external sources are associated with lower annuity amounts.
- Age has a moderate positive correlation with YEARS_EMPLOYED (0.63), suggesting that older individuals generally have more years of employment.
- DAYS_EMPLOYED has a strong negative correlation with AGE (-0.37) and a moderate negative correlation with DAYS_BIRTH (-0.26), indicating that younger individuals tend to be employed for fewer days.
- AMT_INCOME_TOTAL has a moderate positive correlation with AMT_ANNUITY_X (0.76) and AMT_CREDIT_X (0.53), suggesting higher income is associated with higher annuity and credit amounts.
- CNT_CHILDREN has a strong positive correlation with CNT_FAM_MEMBERS (0.88), indicating that families with more children generally have more family members.
- AMT_REQ_CREDIT_BUREAU_YEAR has moderate positive correlations with other AMT_REQ_CREDIT_BUREAU variables (e.g., 0.38 with MONTH, 0.28 with WEEK), showing a consistent pattern in credit bureau requests over different periods.
- Variables related to social circle observations (OBS and DEF) have very low or no significant correlations with other variables, suggesting limited impact on financial metrics.
- DAYS_LAST_PHONE_CHANGE has no significant correlation with most other variables, indicating that phone change history is not related to financial or personal metrics in this dataset.
- The heatmap reveals clusters of variables with strong interrelationships, particularly those related to loan amounts and terms, while social and behavioral variables tend to have weaker

THANK YOU