**CREDIT RISK ANALYTICS (BONDORA BANK)**

**Index**

1. Introduction

* 1.1. Overview of the Project

2. Risk Analytics in Banking and Financial Services

* 2.1. Credit Risk Assessment
* 2.2. Default Prediction
* 2.3. Minimizing Financial Losses
* 2.4. Portfolio and Risk Assessment

3. Problem Statement: Minimizing the Risk of Lending

* 3.1. Risks Associated with Lending
  + 3.1.1. Risk of Approving High-Risk Applicants
  + 3.1.2. Risk of Rejecting Low-Risk Applicants
* 3.2. Objective of the Analysis

4. Data Understanding

* 4.1. application\_data.csv
* 4.2. previous\_application.csv
* 4.3. columns\_description.csv

5. Potential Use Cases

* 5.1. Customer Segmentation
* 5.2. Risk Assessment
* 5.3. Fraud Detection
* 5.4. Product Performance
* 5.5. Customer Lifetime Value (CLTV) Analysis

6. Data Cleaning and Preparation

* 6.1. Handling Missing Data
  + 6.1.1. Analysis of Missing Values
  + 6.1.2. Strategy for Handling Missing Data
* 6.2. Outlier Detection and Handling
  + 6.2.1. Identification of Outliers
  + 6.2.2. Handling of Outliers
* 6.3. Data Transformation
  + 6.3.1. Conversion of Columns to Appropriate Data Types
  + 6.3.2. Manipulation of Categorical and Numerical Data

7. Exploratory Data Analysis (EDA)

* 7.1. Univariate Analysis
  + 7.1.1. Separate Analysis for Target 0 (Non-defaulters) and Target 1 (Defaulters)
    - 7.1.1.1. Categorical Variables
    - 7.1.1.2. Numerical Variables
* 7.2. Bivariate Analysis
  + 7.2.1. Separate Bivariate Analysis for Target 0 and Target 1
    - 7.2.1.1. Categorical Variables
    - 7.2.1.2. Numerical Variables
* 7.3. Data Imbalance Check
  + 7.3.1. Identification of Data Imbalance
  + 7.3.2. Handling Imbalanced Data

8. Correlation Analysis

* 8.1. Correlation Analysis for Target 0 (Non-Defaulters)
  + 8.1.1. Process
  + 8.1.2. Key Findings
* 8.2. Correlation Analysis for Target 1 (Defaulters)
  + 8.2.1. Process
  + 8.2.2. Key Findings
* 8.3. Insights from the Correlation Analysis
  + 8.3.1. Identification of Key Driver Variables
  + 8.3.2. Practical Application

9. Visualizations

* 9.1. Univariate Analysis Visualizations
  + 9.1.1. Categorical Columns
  + 9.1.2. Numerical Columns
* 9.2. Correlation Analysis for Numerical Data with Target Column
  + 9.2.1. Correlation of Non-Defaulters with Numerical Columns
  + 9.2.2. Correlation of Clients with Payment Difficulties
  + 9.2.3. Top 10 Highly Correlated for Non-Defaulter
  + 9.2.4. Top 10 Highly Correlated Dimensions for Defaulters with Payment Difficulties

10. Results

11. Conclusion

**Introduction**

This project focuses on the application of Exploratory Data Analysis (EDA) in a real-world business scenario involving risk analytics in the banking and financial services sector. The goal is to leverage data-driven insights to minimize the risk associated with lending money to customers by analyzing patterns that indicate a client's likelihood of defaulting on a loan. By doing so, the analysis will help ensure that applicants who are capable of repaying their loans are not unfairly rejected, while also protecting the financial institution from potential losses due to high-risk applicants.

**Overview of the Project**

In the lending industry, one of the significant challenges is assessing the creditworthiness of applicants who have insufficient or non-existent credit histories. Such scenarios create a risk for loan providers, as some consumers may exploit the situation and default on their loans. In this project, you are placed in the role of a data analyst working for **Bondora Bank**, a consumer finance company specializing in offering various types of loans to urban customers.

The primary objective is to utilize EDA to analyze patterns in the available data from **Bondora Bank** to ensure that applicants who are capable of repaying the loan are not mistakenly rejected, while those who are likely to default are identified early in the process. The project involves working with a dataset that contains information on loan applications and their outcomes. By examining this data, the goal is to identify key variables that influence whether an applicant is likely to default on their loan.

**Risk Analytics in Banking and Financial Services:**

Risk analytics involves the use of data analysis techniques to identify, measure, and manage risks in banking and financial services. The primary goal of risk analytics is to minimize financial losses while maximizing returns by assessing the potential risk associated with various lending and investment activities.

In the context of lending, risk analytics helps financial institutions understand the likelihood of a borrower defaulting on their loan. This process involves analyzing a variety of customer data, such as credit history, income, employment status, and existing debts, to determine the probability that the borrower will be able to repay the loan.

Key components of risk analytics in banking include:

**Credit Risk Assessment:**

Assessing the creditworthiness of borrowers to determine whether they are likely to default on their loans. This includes evaluating their credit score, credit history, income level, employment status, and other financial factors.

**Default Prediction:**

Identifying patterns and driver variables that indicate the likelihood of a client defaulting on their loan repayments. The analysis focuses on understanding which consumer and loan attributes (such as age, loan amount, employment type, etc.) are strong indicators of default risk.

**Minimizing Financial Losses:**

Using data to make informed decisions about loan approvals, such as whether to approve, reject, or modify the terms of a loan application (e.g., reducing the loan amount or increasing the interest rate for risky applicants). This approach ensures that the institution maintains a balance between minimizing the risk of default and maximizing its business opportunities.

**Portfolio and Risk Assessment:**

Continuously analyzing the risk profile of the entire loan portfolio to ensure that the bank's exposure to high-risk loans is within acceptable limits. This helps in adjusting the lending strategies and managing potential losses.

**Problem Statement: Minimizing the Risk of Lending**

Financial institutions, particularly loan providers, face significant challenges when determining whether to approve or reject loan applications due to potential risks associated with lending. One of the key issues is assessing the risk of lending to applicants who may have insufficient or non-existent credit histories. In such cases, it becomes difficult to predict their likelihood of repaying the loan.

To address this problem, a consumer finance company that specializes in providing various types of loans to urban customers needs to analyze the patterns in its data to ensure that only applicants capable of repaying their loans are approved, while those likely to default are identified and managed appropriately.

**Risks Associated with Lending:**

1. **Risk of Approving High-Risk Applicants:**

Approving loans to applicants who are not likely to repay (defaulters) poses a significant financial risk to the company. This can result in substantial financial losses due to unpaid loans.

1. **Risk of Rejecting Low-Risk Applicants:**

Rejecting loans to applicants who are likely to repay leads to a loss of business opportunities and potential revenue. This situation can occur when creditworthy applicants are mistakenly categorized as high-risk due to inadequate or incomplete information.

**Objective of the Analysis:**

The objective is to identify patterns and driving factors (driver variables) that indicate a client’s difficulty in making loan payments. This analysis will help in:

* Deciding whether to approve or reject a loan application.
* Adjusting loan terms such as the amount or interest rate based on the applicant's risk profile.
* Ensuring that the company minimizes its risk exposure by approving loans only to those applicants who are most likely to repay.

**Data Understanding**

**1. application\_data.csv**

* **Purpose:** Contains information about clients at the time of application, focusing on factors related to payment difficulties.

**2. previous\_application.csv**

* **Purpose:** Contains information about the client's previous loan applications.

**3. columns\_description.csv**

* **Purpose:** Provides descriptions and definitions for the variables in the datasets.

**Potential Use Cases**

* **Customer Segmentation:** Analyzing customer data based on product usage, transaction history, payment behavior, and other factors to identify different customer segments.
* **Risk Assessment:** Evaluating customer risk based on factors such as payment history, rejection rates, account activity, and product usage.
* **Fraud Detection:** Identifying patterns of fraudulent activity by analyzing transaction data, comparing it to historical trends, and considering factors like rejection rates and account status.
* **Product Performance:** Assessing the performance of different products or services based on sales, customer satisfaction, and other metrics.
* **Customer Lifetime Value (CLTV) Analysis:** Calculating the expected long-term value of customers based on their behavior and characteristics.

**Data Cleaning and Preparation**

**Handling Missing Data:**

1. Analysis of Missing Values:

The analysis identified the percentage of missing values in each column of the dataset.

Columns with more than 50% missing values were identified and subsequently dropped from the dataset. This threshold was chosen as these columns have too many missing values to be reliably imputed.

2. Strategy for Handling Missing Data:

Columns with more than 50% missing values were removed to avoid bias or inaccuracies.

For the remaining columns with missing values, the data was imputed using appropriate methods:

Mode Imputation: For categorical variables and numerical variables with distinct modes (e.g., `OCCUPATION\_TYPE`, `AMT\_GOODS\_PRICE`), missing values were replaced with the mode of the column.

Zero Imputation: Some columns (e.g., `AMT\_REQ\_CREDIT\_BUREAU\_MON`, `AMT\_REQ\_CREDIT\_BUREAU\_WEEK`, etc.) had a majority of values as zero. Missing values in these columns were imputed with zero.

**Outlier Detection and Handling:**

1. Identification of Outliers:

Outliers were identified using the Interquartile Range (IQR) method. The upper and lower bounds were calculated to detect potential outliers for various columns (e.g., `AMT\_REQ\_CREDIT\_BUREAU\_MON`, `AMT\_ANNUITY`, etc.).

2. Handling of Outliers:

Outliers were not removed, and the decision was made to retain them. This choice is based on the nature of the data, where the outliers may represent genuine variations in the population rather than errors or anomalies.

For certain columns, despite identifying outliers, the imputation was done using the mode or zero, as these outliers did not significantly distort the overall data distribution.

**Data Transformation:**

1. Conversion of Columns to Appropriate Data Types:

Columns were reviewed and converted to appropriate data types. For example, categorical variables were correctly identified and converted to object types, while numerical columns were retained as float or integer types.

2. Manipulation of Categorical and Numerical Data:

Categorical Data:

Columns such as `NAME\_CONTRACT\_TYPE`, `FLAG\_OWN\_CAR`, and others were identified as categorical. Univariate analysis was conducted to observe how these variables behave across different target classes.

Numerical Data:

Continuous numerical columns were analyzed for their distribution and relationship with the target variable. The correlation between numerical features and the target variable was visualized using heatmaps to understand the linear relationships.

**Exploratory Data Analysis (EDA)**

**Univariate Analysis:**

Univariate analysis involves examining the distribution of individual variables to understand their properties. This includes visualizing the data to observe patterns, central tendency, and variability.

1. **Separate Analysis for Target 0 (Non-defaulters) and Target 1 (Defaulters):**

**Categorical Variables:**

* The distribution of categorical variables such as NAME\_CONTRACT\_TYPE, FLAG\_OWN\_CAR, FLAG\_OWN\_REALTY, etc., was analyzed separately for non-defaulters (Target 0) and defaulters (Target 1).
* Bar plots were used to visualize how these variables are distributed across the two target classes.
* For example, the ownership of cars (FLAG\_OWN\_CAR) and real estate (FLAG\_OWN\_REALTY) was compared between non-defaulters and defaulters, highlighting any noticeable differences.

**Numerical Variables:**

* Numerical variables such as AMT\_GOODS\_PRICE, DAYS\_LAST\_PHONE\_CHANGE, AMT\_INCOME\_TOTAL, etc., were analyzed using box plots and distribution plots.
* Separate box plots were created for Target 0 and Target 1 to observe the spread and potential outliers within each group.
* The analysis revealed how these variables differ between non-defaulters and defaulters, providing insights into characteristics that may be associated with loan defaulting behavior.

**Bivariate Analysis:**

Bivariate analysis examines the relationship between two variables, which helps in understanding dependencies or associations that might exist between them.

1. **Separate Bivariate Analysis for Target 0 and Target 1:**

**Categorical Variables:**

* The relationship between pairs of categorical variables was explored for both target classes. For instance, the relationship between NAME\_INCOME\_TYPE and NAME\_EDUCATION\_TYPE was visualized to see how income types are distributed across different education levels for non-defaulters and defaulters.
* Cross-tabulations and stacked bar charts were likely used to show the interaction between these categorical variables.

**Numerical Variables:**

* Correlation matrices and heatmaps were used to explore the relationships between numerical variables for non-defaulters and defaulters separately.
* This helped in identifying variables that are strongly correlated with each other within each target class, providing insights into potential predictors of loan defaulting.

**Data Imbalance Check:**

1. **Identification of Data Imbalance:**

* The target variable was checked for class imbalance. The analysis showed a significant imbalance, with the majority of the data belonging to non-defaulters (Target 0) and a smaller portion belonging to defaulters (Target 1).
* The distribution was visualized using pie charts, where the percentage of non-defaulters was much higher than that of defaulters.

1. **Handling Imbalanced Data:**

* The document suggests recognizing the imbalance as a potential issue that could affect model performance. No specific method (e.g., oversampling, under sampling, or using a balanced dataset) was detailed in the provided excerpts, but the imbalance was clearly identified.
* The visualization of data distribution and the pie chart depicting the class proportions were used to illustrate the extent of the imbalance.

**Correlation Analysis**

* Correlation analysis involves examining the strength and direction of linear relationships between pairs of numerical variables. In the context of this dataset, the goal was to identify which variables are most strongly associated with each other and how these associations differ between non-defaulters (Target 0) and defaulters (Target 1).

**1. Correlation Analysis for Target 0 (Non-Defaulters):**

* **Process:**

A correlation matrix was created for non-defaulters, focusing on numerical variables. The matrix provides a comprehensive view of how variables relate to each other within the non-defaulter group.

The correlations were then filtered to identify the top 10 pairs of variables with the highest absolute correlation values.

* **Key Findings:**

The top 10 correlated variable pairs for non-defaulters were identified, which might include strong relationships between income-related variables, credit amounts, and loan terms.

For example, variables like AMT\_CREDIT (credit amount) and AMT\_GOODS\_PRICE (goods price) might show a strong positive correlation, indicating that higher credit amounts are often associated with higher goods prices for non-defaulters.

**2. Correlation Analysis for Target 1 (Defaulters):**

* **Process:**

Similarly, a correlation matrix was created for defaulters, and the top 10 correlated pairs of variables were identified.

This analysis helps in understanding if different variables drive the defaulting behavior compared to non-defaulting behavior.

* **Key Findings:**

The top 10 correlated pairs for defaulters may differ from those of non-defaulters, indicating that certain financial behaviors or attributes are more pronounced in the group of defaulters.

For instance, variables like DAYS\_EMPLOYED (employment duration) might show a different pattern of correlation with loan amounts or income for defaulters compared to non-defaulters, suggesting different risk profiles.

**3. Insights from the Correlation Analysis:**

* **Identification of Key Driver Variables:**

The correlation analysis helps in pinpointing key variables that are strongly associated with loan default. These might include variables related to credit amounts, income levels, employment duration, and other financial factors.

By comparing the correlation patterns between non-defaulters and defaulters, the analysis can reveal which variables are likely to be key drivers of default risk.

* **Practical Application:**

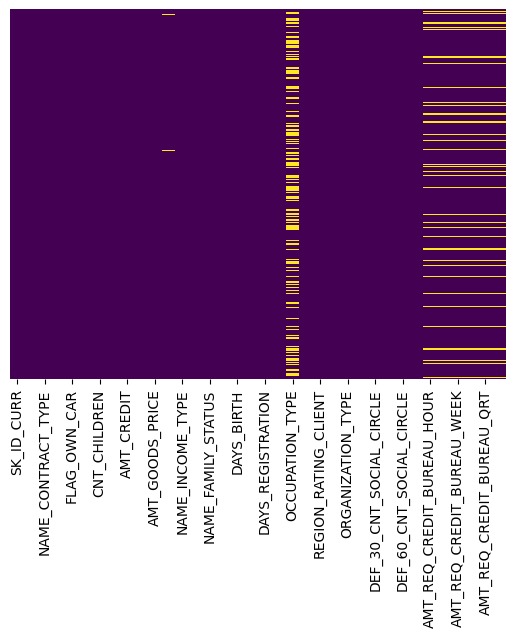
The identified correlations provide a basis for selecting features that may be important in predictive modeling. Variables that show high correlation with each other or with the target variable (default status) can be considered significant in predicting loan default risk.

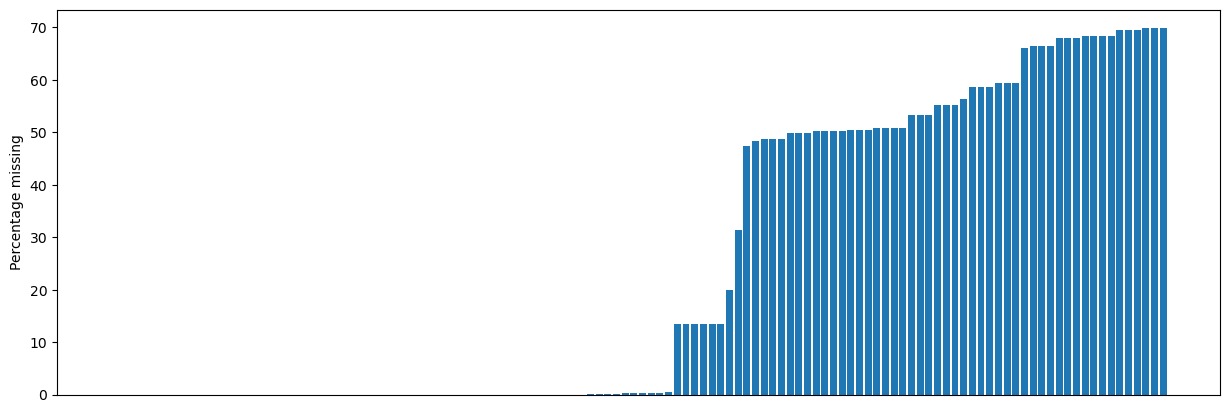
For example, if AMT\_CREDIT and AMT\_ANNUITY (annuity amount) show strong correlations with default status in defaulters, they might be crucial variables to include in a risk assessment model.

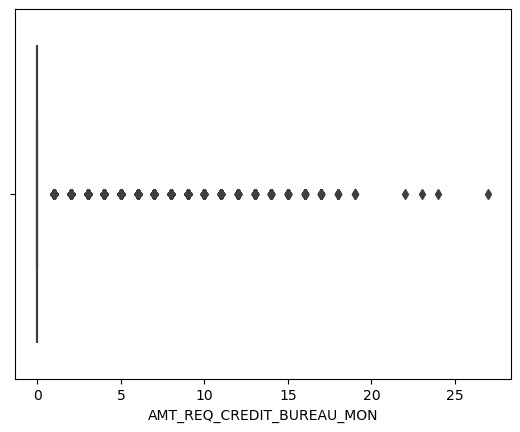
**Visualizations**

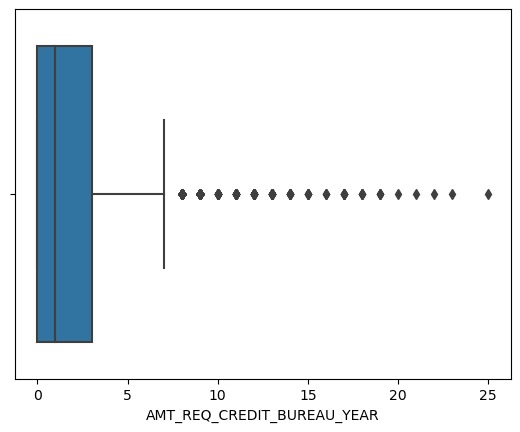
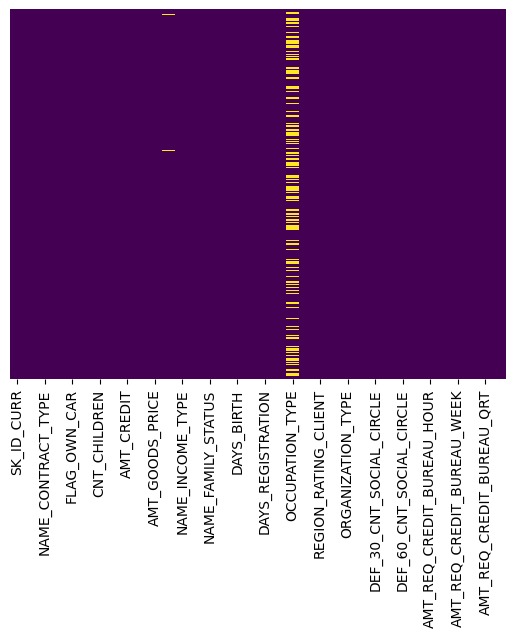
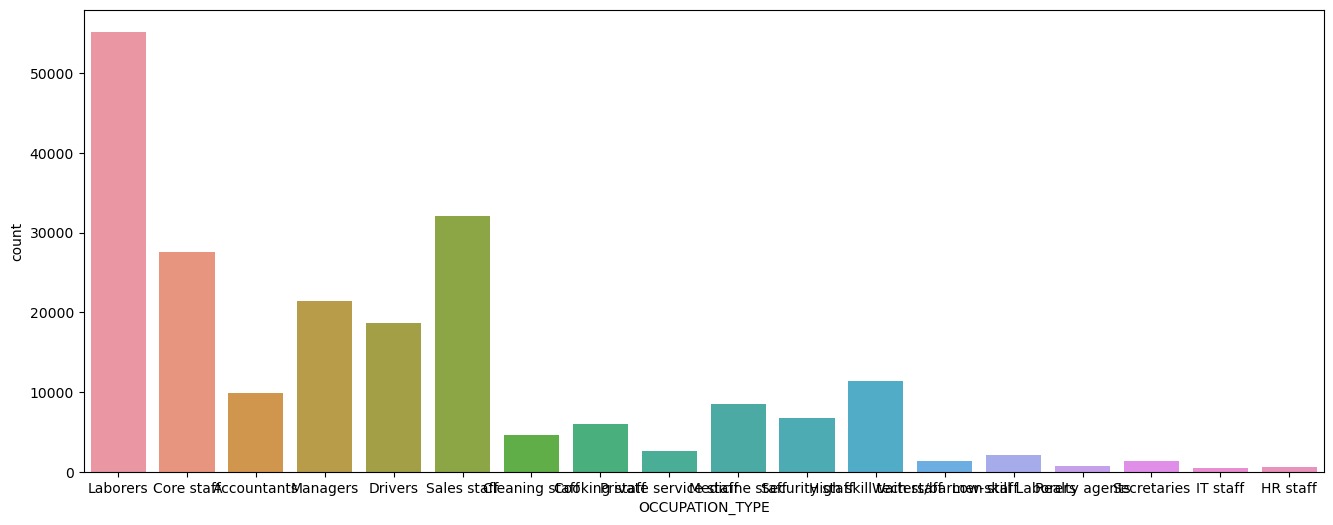
**Univariate Analysis Visualizations:**

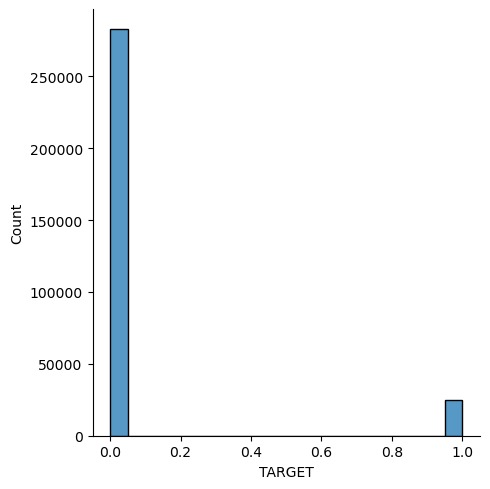
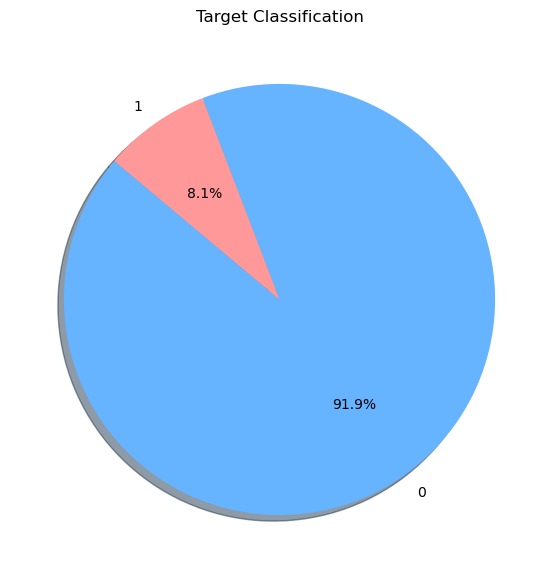
Categorical Columns:

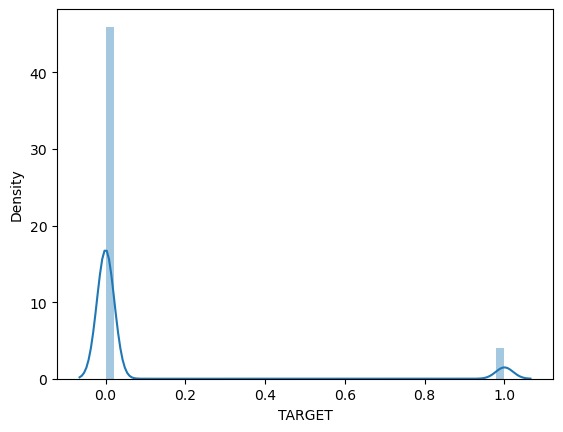




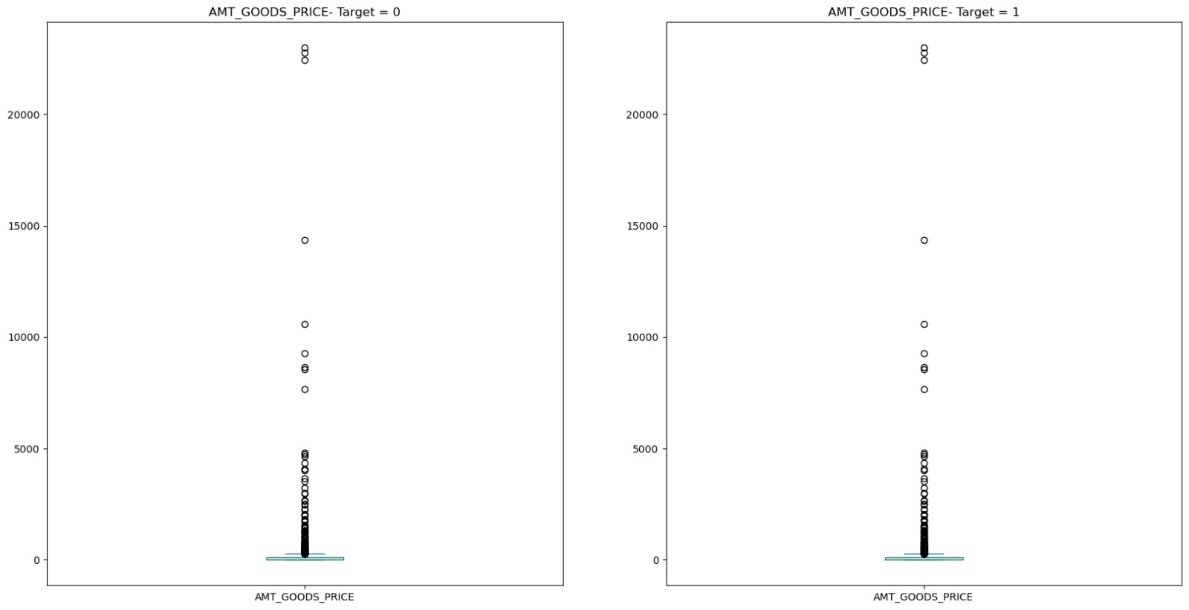


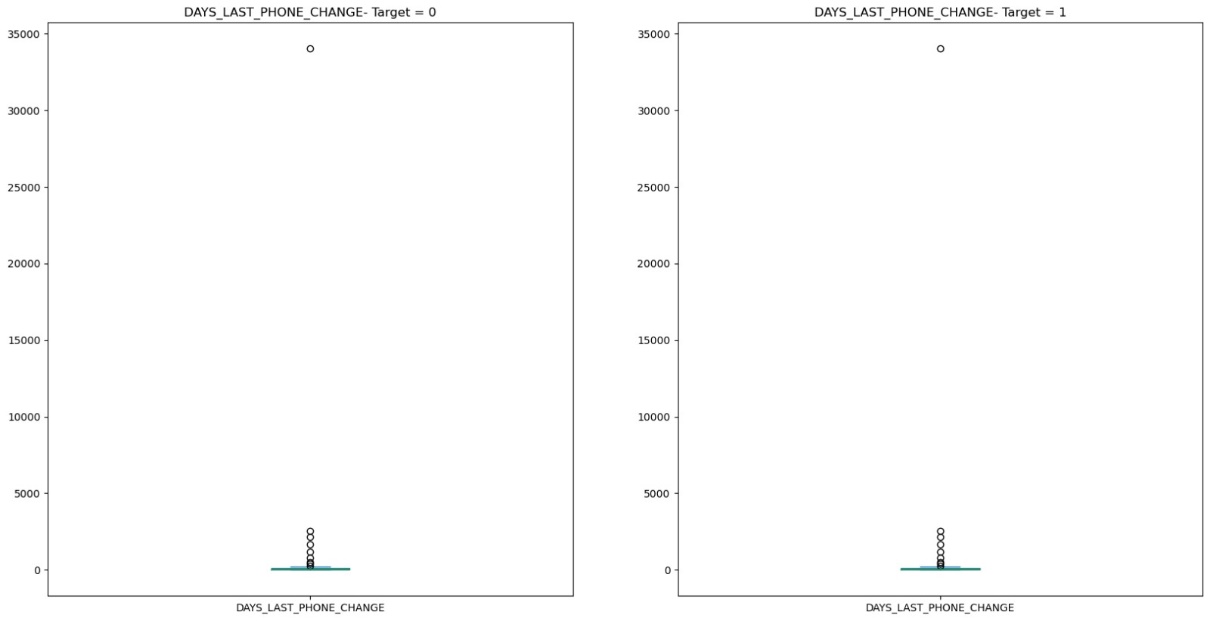
   


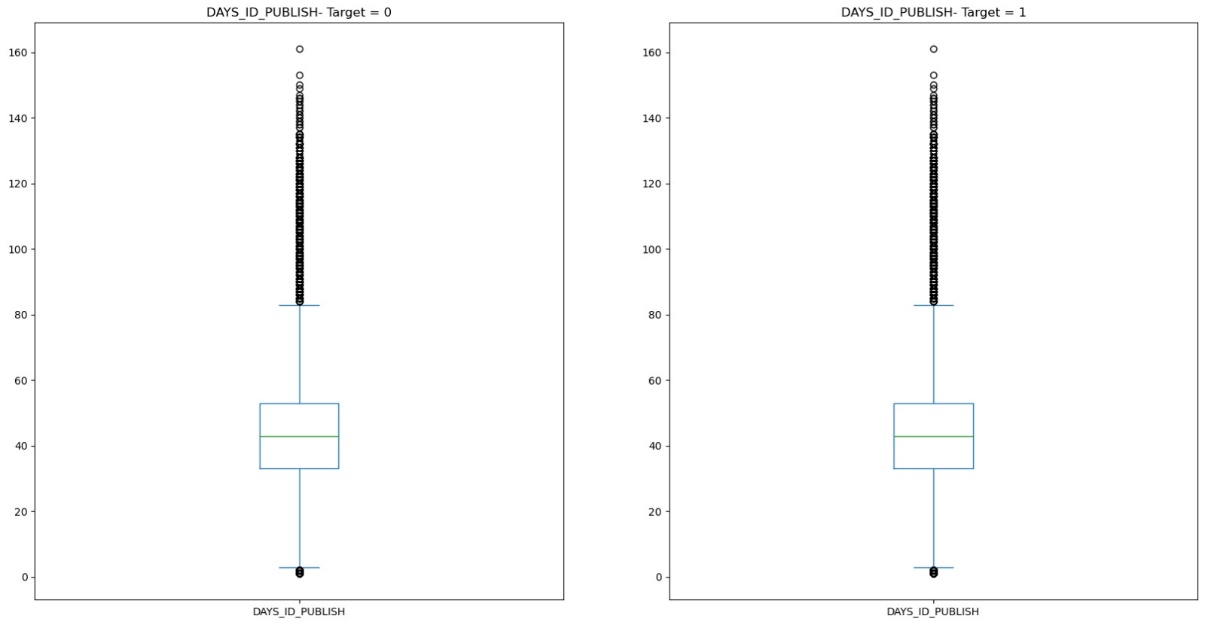


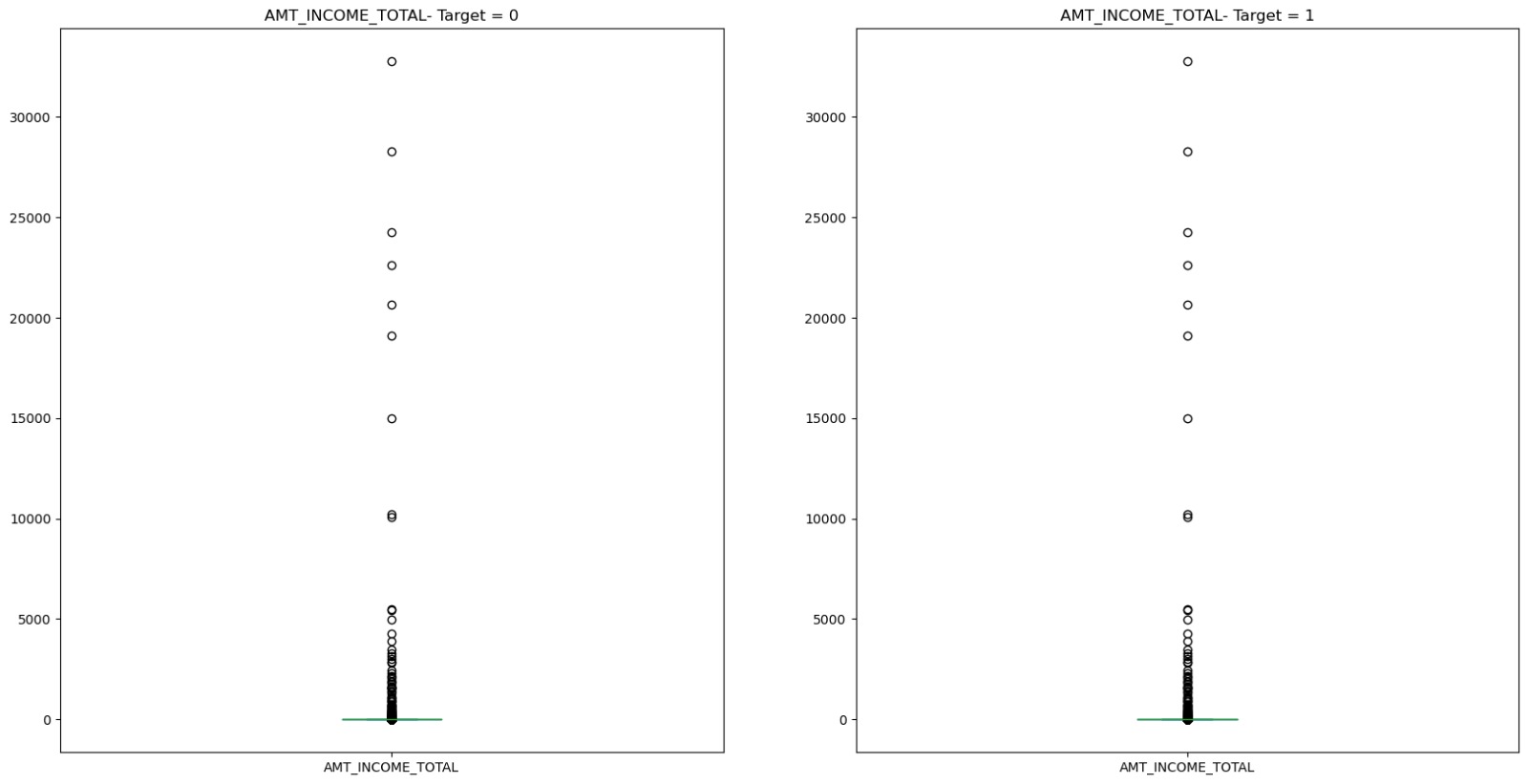


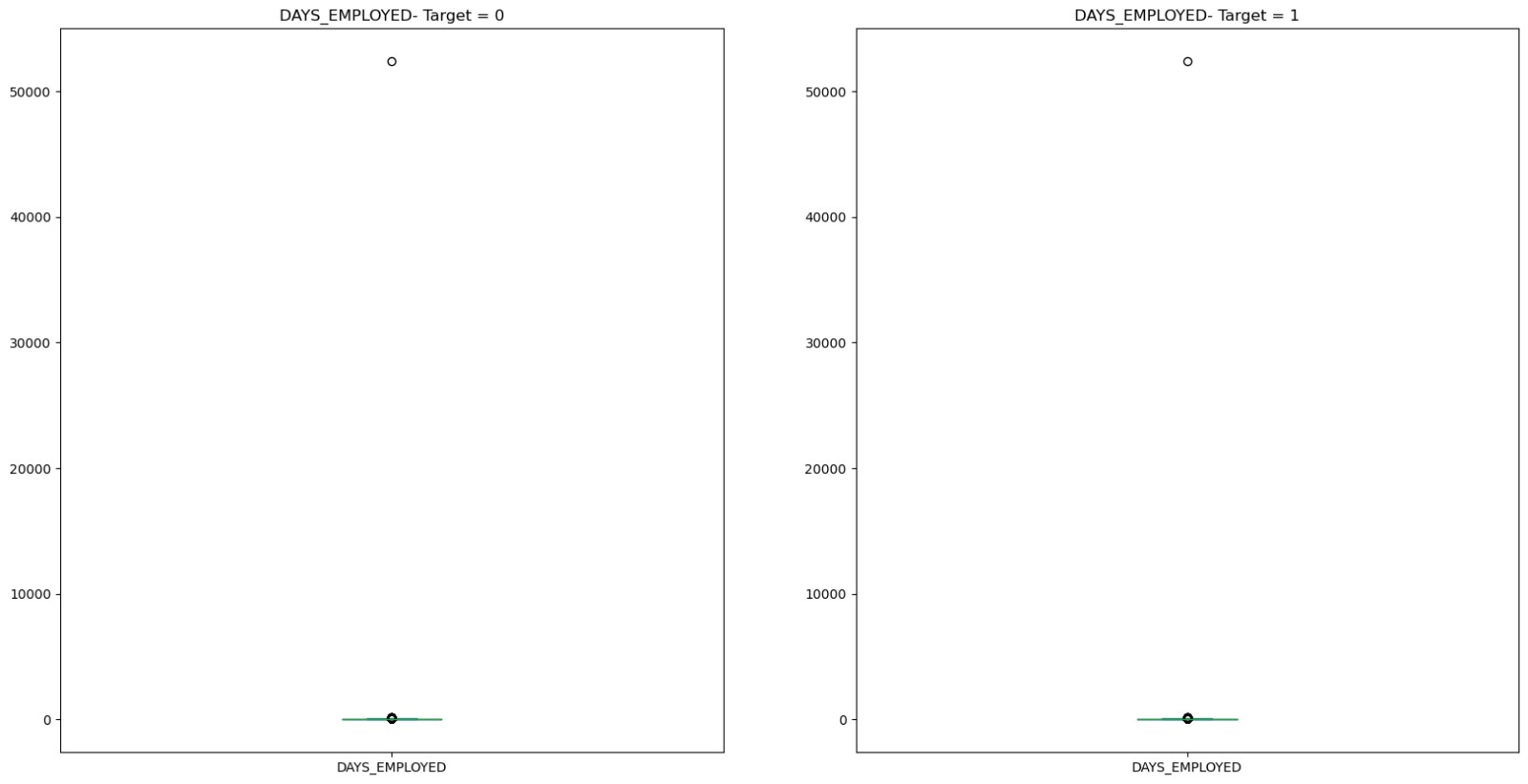
Numerical Columns:

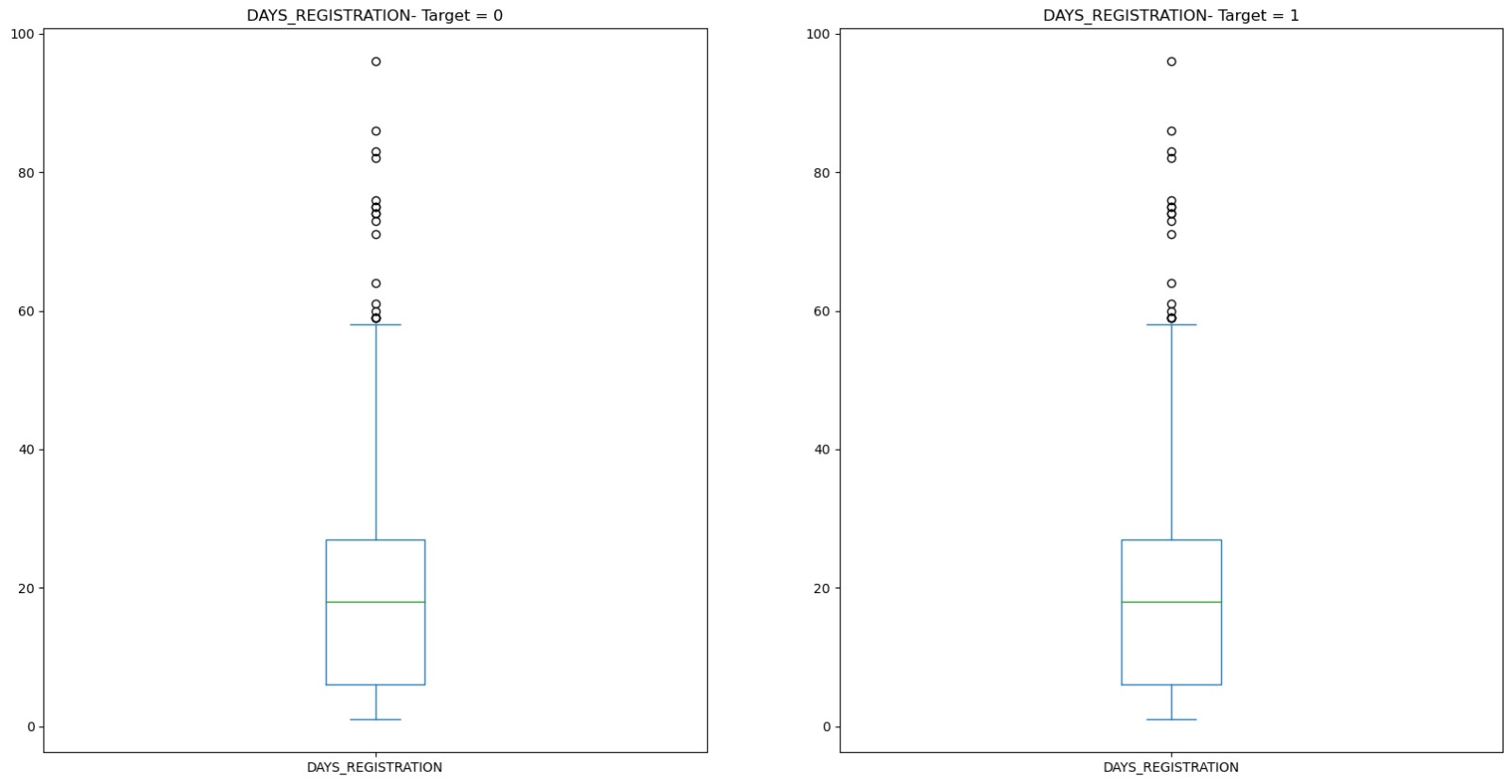
****

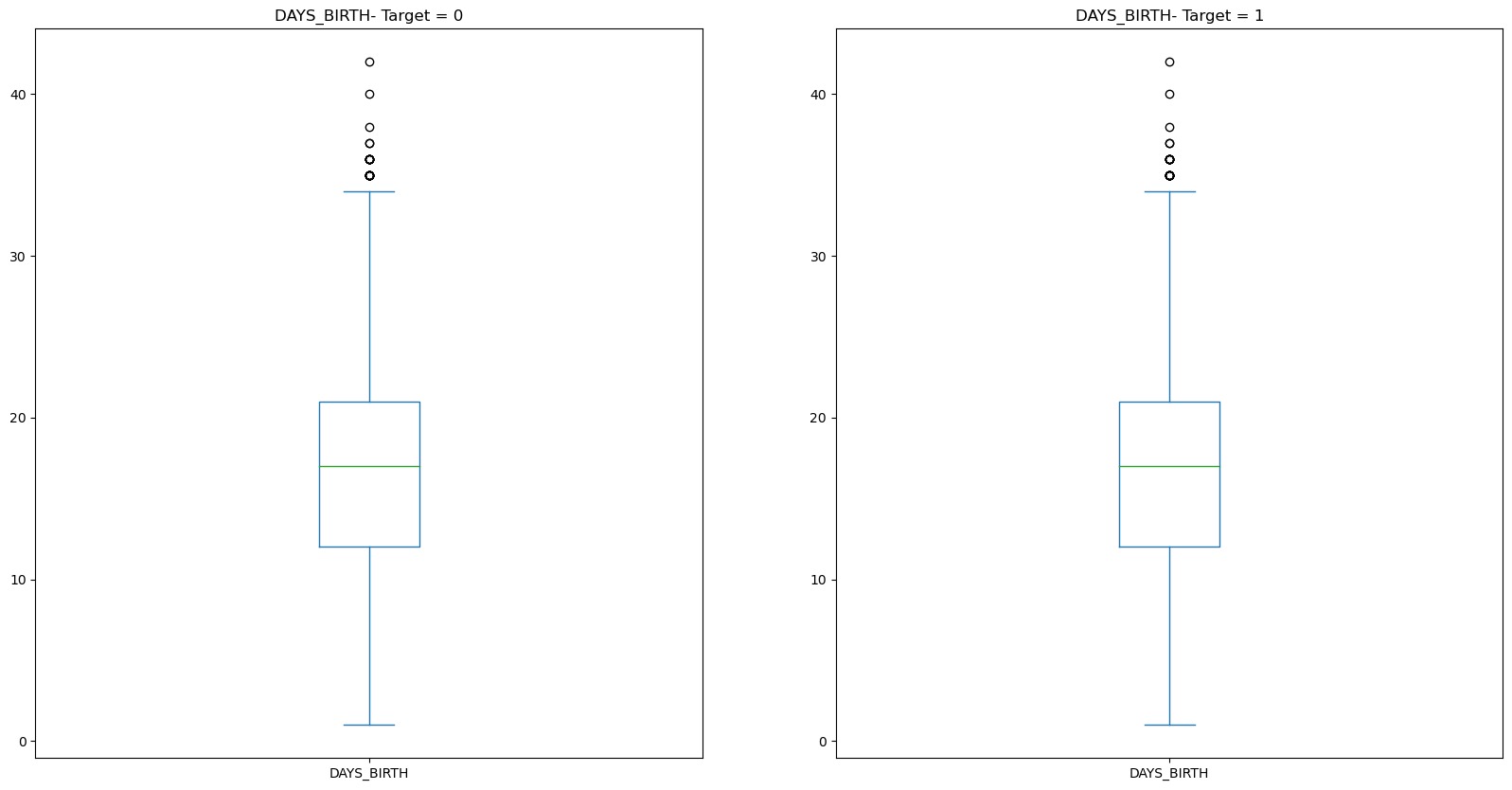
****

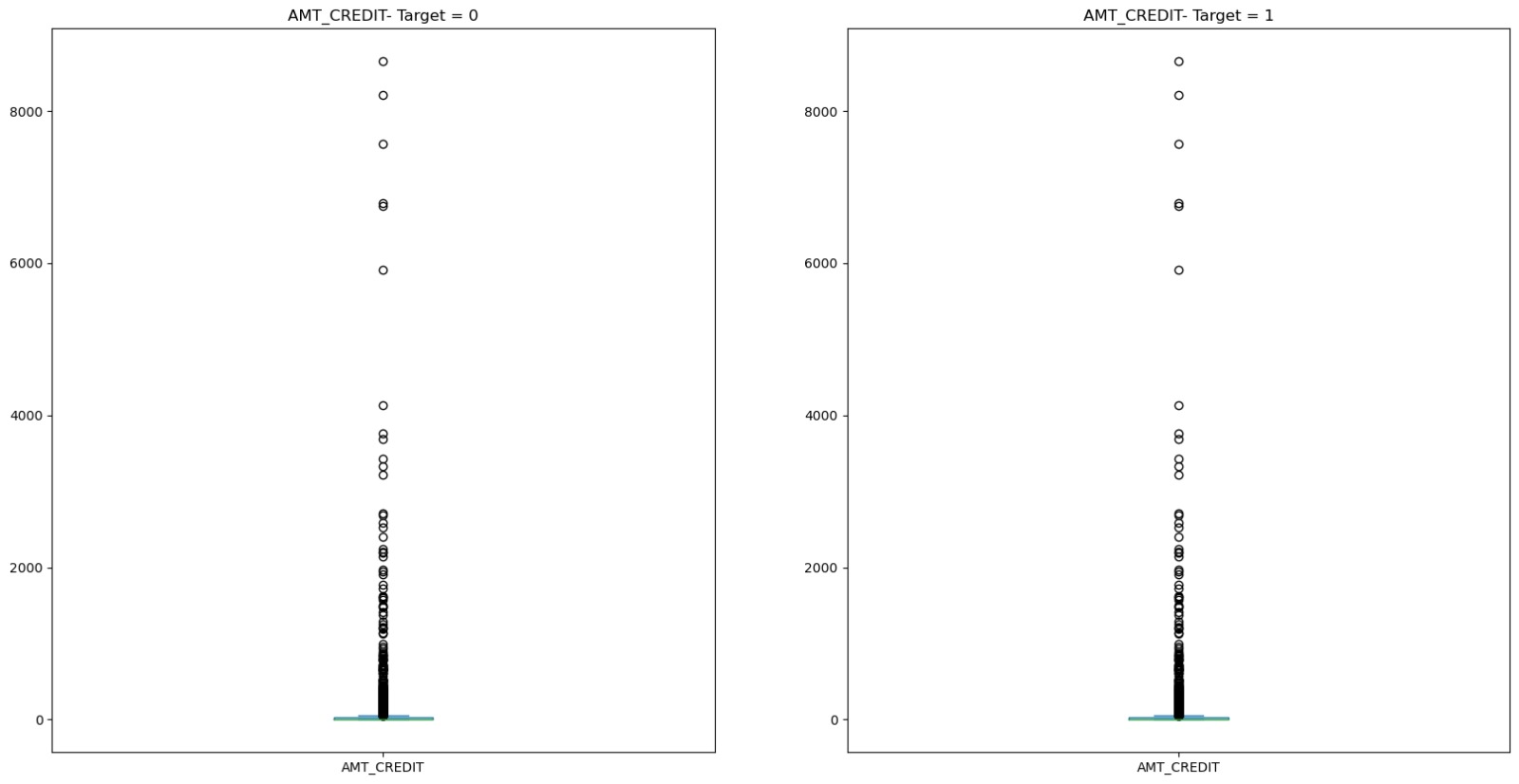
****

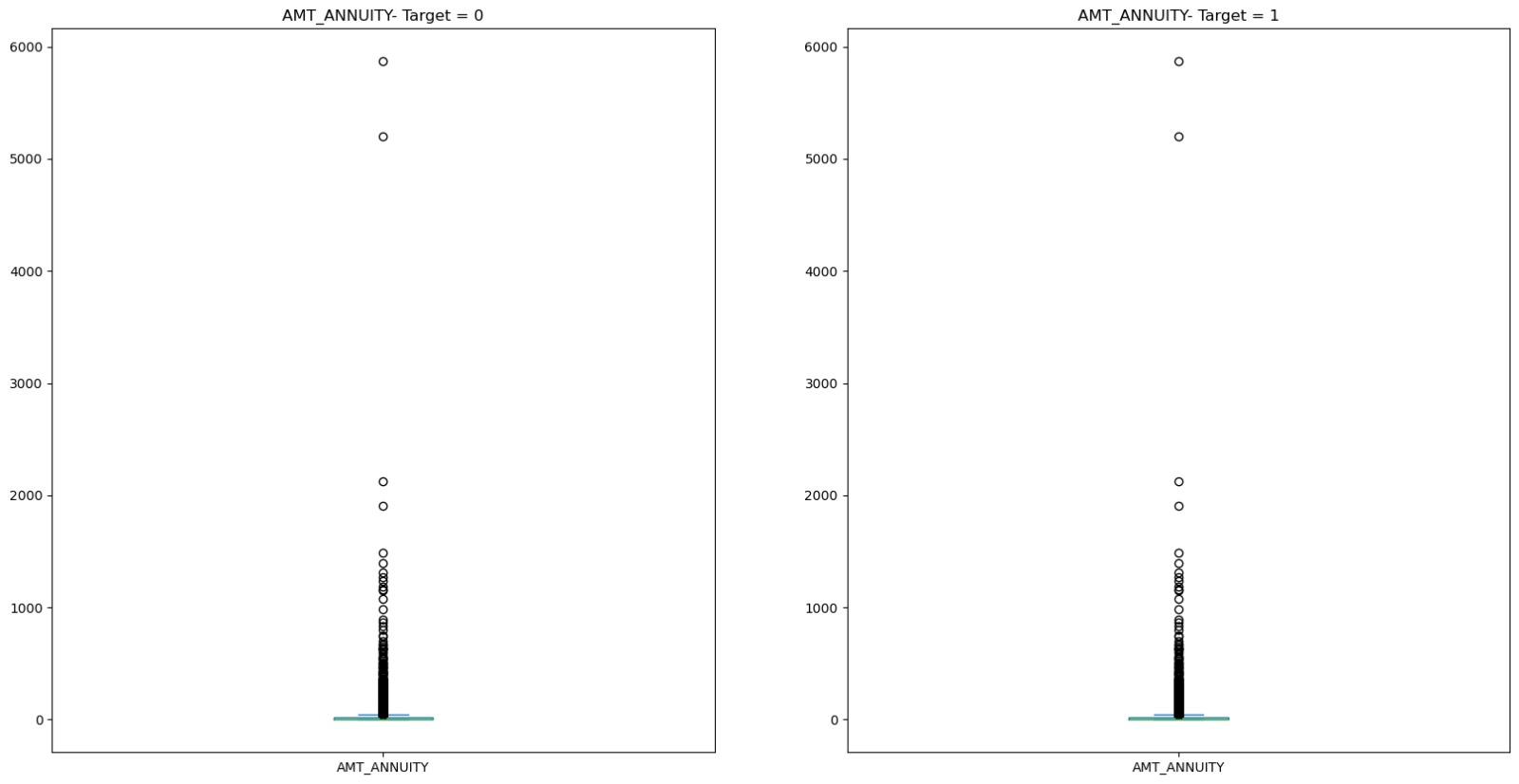
****

****

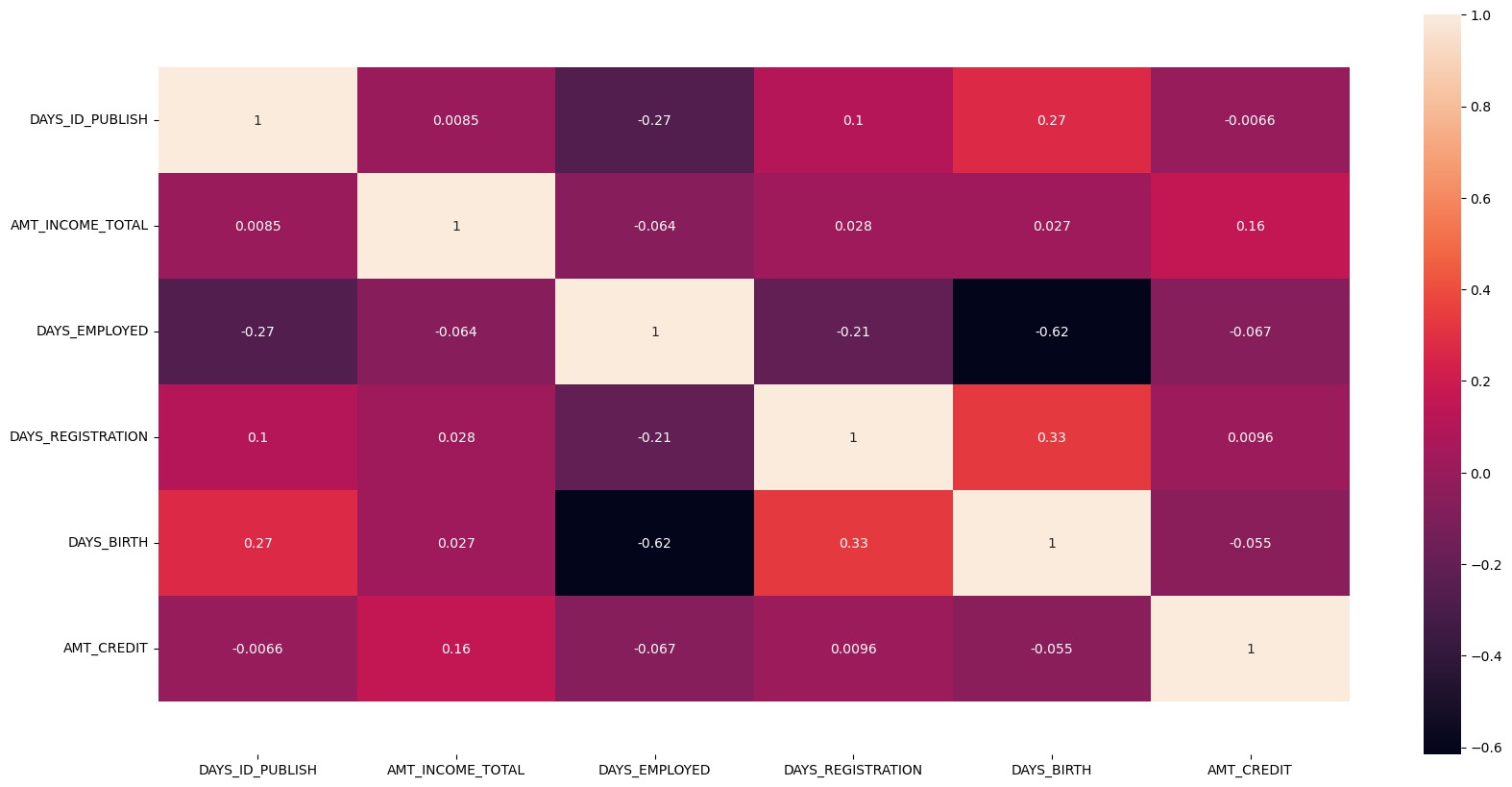
****

****

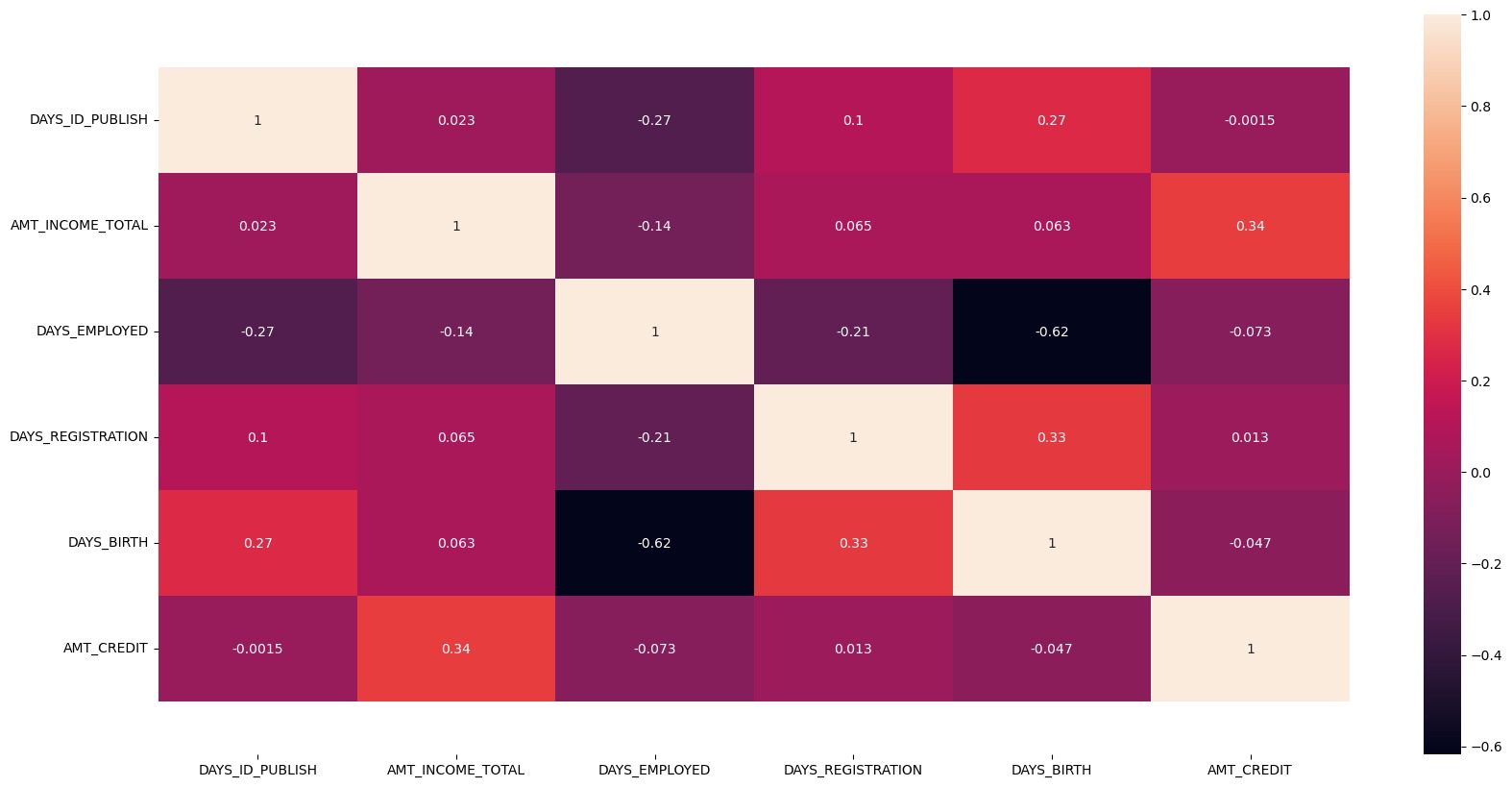
****

****

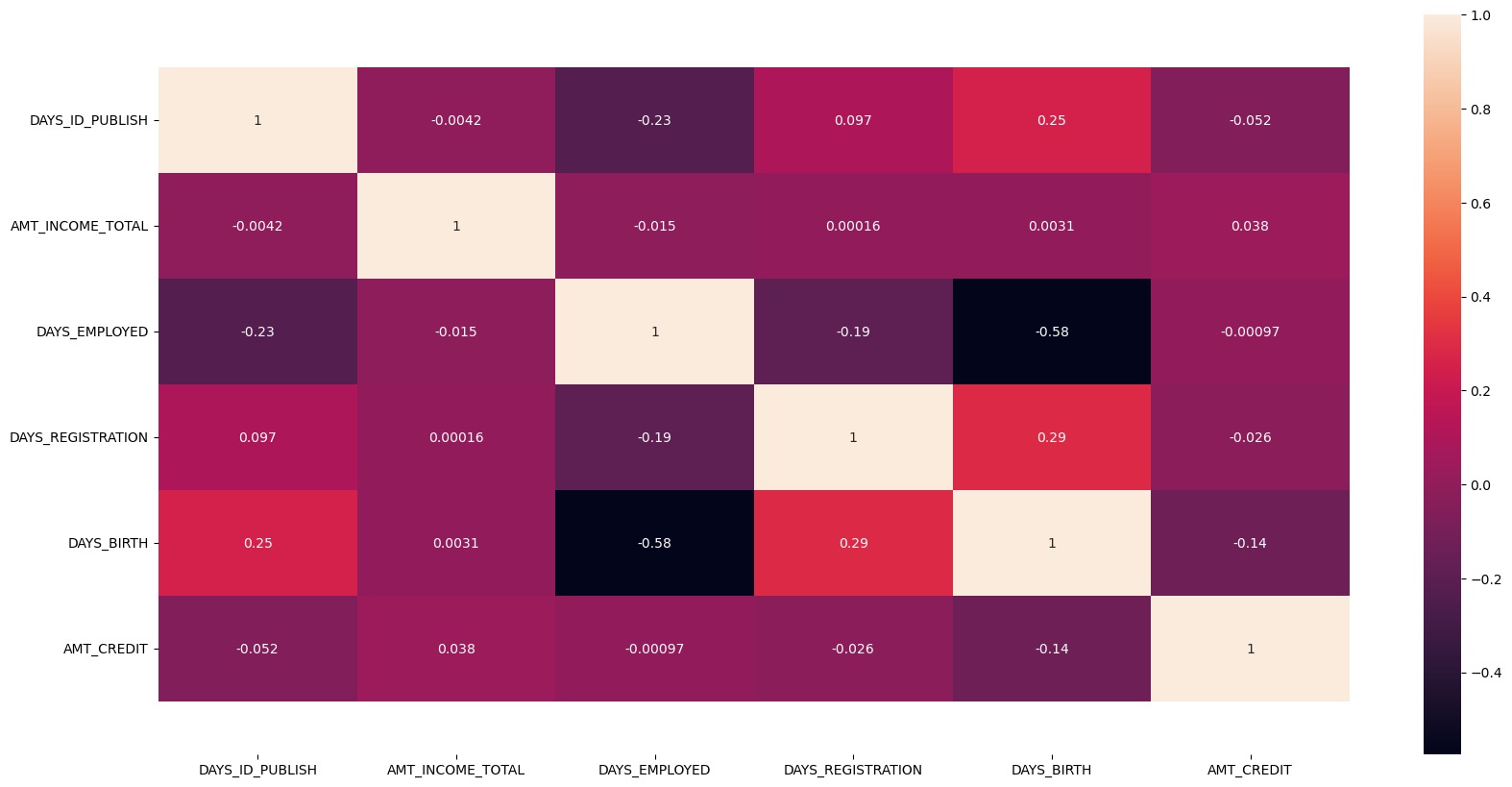
**Correlation analysis for Numerical data with Target column:**



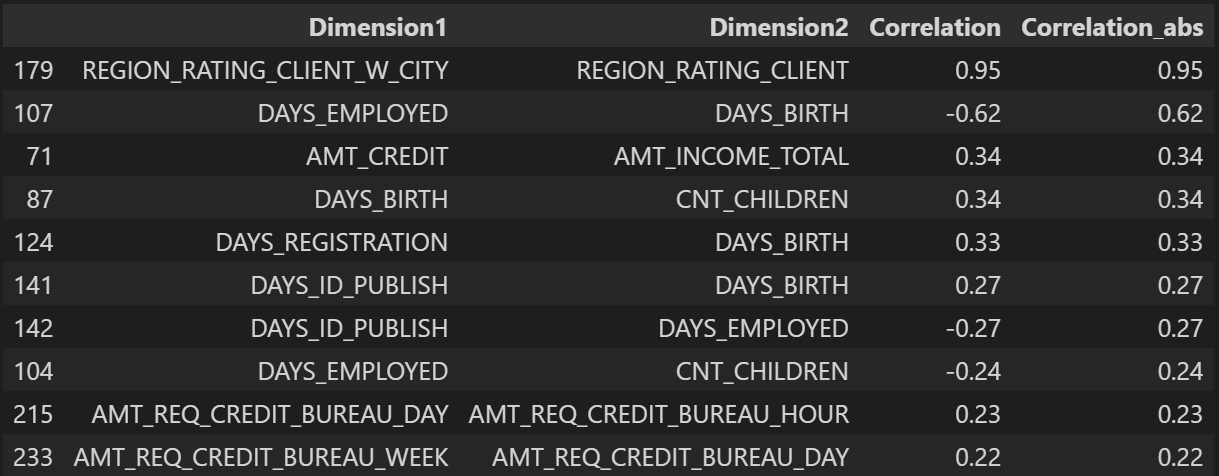
**Correlation of non-defaulters with numerical columns:**



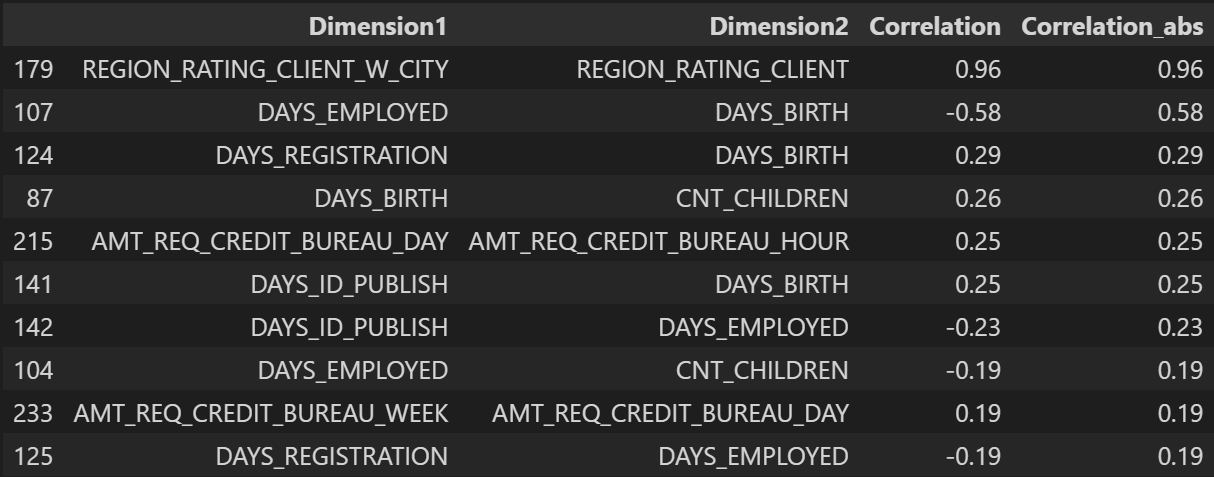
**Correlation of clients with payment difficulties:**

****

**Top 10 Highly Correlated for Non-Defaulter:**

****

**Top 10 Highly Correlated Dimensions for Defaulters with Payment Difficulties:**

****

**Results**

The analysis of the loan application data identified key driver variables associated with loan default, such as `AMT\_CREDIT`, `AMT\_ANNUITY`, and `DAYS\_EMPLOYED`, which show strong correlations with payment difficulties. The exploration revealed that higher credit amounts and annuity payments, along with shorter employment durations, are significant indicators of potential default. The data also showed a significant imbalance, with a majority of non-defaulters compared to defaulters, emphasizing the need for tailored risk assessment strategies. By focusing on these key variables, the company can better predict loan defaults and take preemptive actions, such as adjusting loan terms or interest rates, to minimize financial risk**.**

**Conclusion**

The exploratory data analysis (EDA) conducted on the loan application dataset provided valuable insights into the factors that influence loan default. Key variables such as `AMT\_CREDIT`, `AMT\_ANNUITY`, and `DAYS\_EMPLOYED` emerged as significant predictors of payment difficulties, indicating that larger loan amounts, higher annuity payments, and shorter employment durations are associated with a higher risk of default. The analysis also highlighted a significant imbalance in the dataset, with a much larger proportion of non-defaulters compared to defaulters. This imbalance underscores the importance of implementing tailored risk assessment strategies to ensure that loans are only approved for applicants who are likely to repay, while also managing the risk of potential defaults. By leveraging these insights, the company can refine its lending practices, optimize its risk management, and ultimately improve its financial stability by minimizing losses associated with loan defaults.