## Data Preprocessing

Based on previous considerations, we selected variables below to later use them in our model.

Table X: Variable Encoding Table

|  |  |  |
| --- | --- | --- |
| **Var\_Title** | **Var\_Description** | **Field\_Content** |
| PAYMENT\_DAY | Day of the month for bill payment, chosen by t... | 1,5,10,15,20,25 |
| APPLICATION\_SUBMISSION\_TYPE | Indicates if the application was submitted via... | Web, Carga |
| POSTAL\_ADDRESS\_TYPE | Indicates if the address for posting is the ho... | 1.2 |
| MARITAL\_STATUS | Encoding not informed | 1,2,3,4,5,6,7 |
| QUANT\_DEPENDANTS | NaN | 0, 1, 2, ... |
| CITY\_OF\_BIRTH | NaN | NaN |
| RESIDENCIAL\_STATE | State of residence | NaN |
| RESIDENCIAL\_BOROUGH | Borough of residence | NaN |
| FLAG\_RESIDENCIAL\_PHONE | Indicates if the applicant possesses a home phone | Y,N |
| MONTHS\_IN\_RESIDENCE | Time in the current residence in months | 1,2,... , NULL |
| FLAG\_MOBILE\_PHONE | Indicates if the applicant possesses a mobile ... | Y,N |
| FLAG\_EMAIL | Indicates if the applicant possesses an e-mail... | 0.1 |
| PERSONAL\_MONTHLY\_INCOME | Applicant's personal regular monthly income in... | NaN |
| OTHER\_INCOMES | Applicant's other incomes monthly averaged in ... | NaN |
| FLAG\_VISA | Flag indicating if the applicant is a VISA cre... | 0.1 |
| FLAG\_MASTERCARD | Flag indicating if the applicant is a MASTERCA... | 0.1 |
| FLAG\_DINERS | Flag indicating if the applicant is a SINERS c... | 0.1 |
| FLAG\_AMERICAN\_EXPRESS | Flag indicating if the applicant is an AMERICA... | 0.1 |
| QUANT\_BANKING\_ACCOUNTS | NaN | 0,1,2 |
| QUANT\_SPECIAL\_BANKING\_ACCOUNTS | NaN | 0,1,2 |
| PERSONAL\_ASSETS\_VALUE | Total value of the personal possessions such a... | NaN |
| QUANT\_CARS | Quantity of cars the applicant possesses | NaN |
| COMPANY | If the applicant has supplied the name of the ... | Y,N |
| FLAG\_PROFESSIONAL\_PHONE | Indicates if the professional phone number was... | Y,N |
| MONTHS\_IN\_THE\_JOB | Time in the current job in months | NaN |
| FLAG\_ACSP\_RECORD | Flag indicating if the applicant has any previ... | Y, N |
| AGE | Applicant's age at the moment of submission | NaN |

## Missing Value Handling

### Substituting with Mode:

PAYMENT\_DAY: The outlier value of -99999, presumably a placeholder for missing data, was replaced with the mode of PAYMENT\_DAY.

MONTHS\_IN\_RESIDENCE: Zero values, potentially indicating data collection errors, were replaced with the median (6 months) for Brazilian nationals, assuming they're likely not new immigrants. For non-Brazilians, zeros were retained under the assumption that they represent new migrants.

### Categorical Conversion and Substitution:

APPLICATION\_SUBMISSION\_TYPE and MARITAL\_STATUS: Zero values in both variables, likely indicating missing information, were replaced with a distinct category, "Unknown" for APPLICATION\_SUBMISSION\_TYPE, and the most common marital status for MARITAL\_STATUS. This approach maintains data integrity and avoids misinterpretation.

RESIDENCIAL\_STATE: To address data sparsity and the curse of dimensionality, values were categorized into five regions, transforming RESIDENCIAL\_STATE into a new variable, RESIDENCIAL\_REGION, based on geographical mapping (Table X1).

Table X1: Geographical Mapping Table

|  |  |
| --- | --- |
| **Direction** | **Region Codes** |
| Southeast | SP, RJ, MG, ES |
| South | RS, SC, PR |
| Northeast | BA, CE, PE, RN, AL, PB, SE, PI, MA |
| Central-West | GO, MT, MS, DF |
| North | PA, AP, AM, RO, RR, TO, AC |

## Redundant Information

We treated RESIDENCIAL\_BOROUGH and CITY\_OF\_BIRTH as redundant information as RESIDENCIAL\_REGION since they both indicates the residential area of the applicants.

## Dropping Variables due to Non-Informativeness or Legal Considerations

FLAG\_MOBILE\_PHONE, FLAG\_ACSP\_RECORD, and PERSONAL\_ASSETS\_VALUE were dropped due to all values being zero, providing no informative variation (Neptune.ai, 2023). MONTHS\_IN\_THE\_JOB was dropped because of its highly right-skewed distribution with a majority of values at zero, potentially leading to data sparsity issues. NATIONALITY was removed in compliance with the Equal Credit Opportunity Act (ECOA) (Federal Reserve Board. 2007), which prohibits discrimination based on national origin. Observations in AGE below 18 were dropped, assuming that individuals under 18 are not legally considered in credit scoring.

Neptune.ai. (2023). A Comprehensive Guide to Data Preprocessing. Available at: <https://neptune.ai/blog/data-preprocessing-guide> [Accessed 28 Nov. 2023].

Federal Reserve Board. (2007). *Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit*. Available at: <https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/demographics.htm> [Accessed 28 November 2023].

## Retaining or Modifying Selective Predictors

For QUANT\_SPECIAL\_BANKING\_ACCOUNTS and QUANT\_BANKING\_ACCOUNTS, due to identical values, only one of the predictors was retained to avoid redundancy. NUMBER OF DEPENDENT was kept, as a large number of zeros (indicating no dependents) is deemed reasonable and provides valuable information.

## Outlier Handling

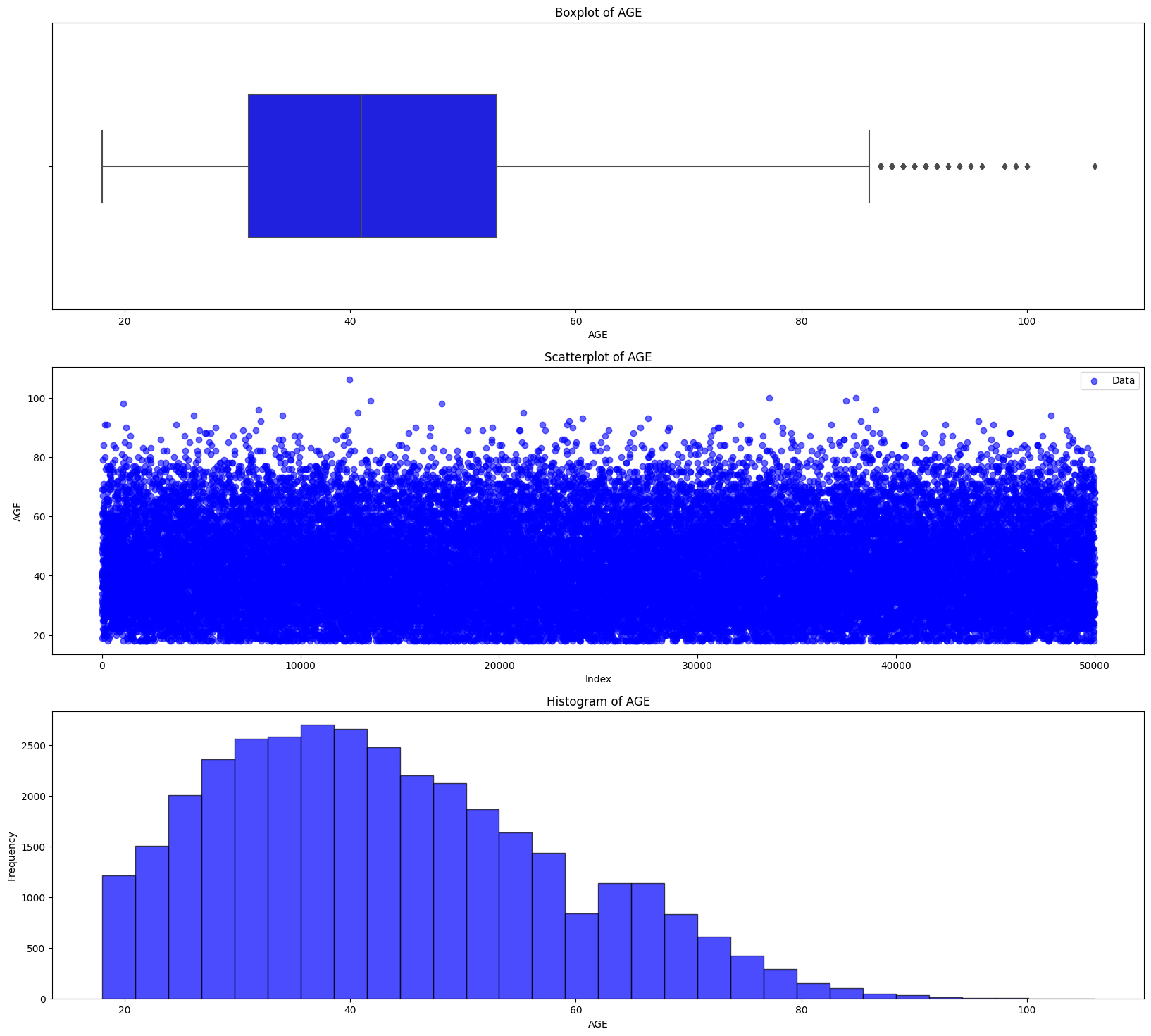
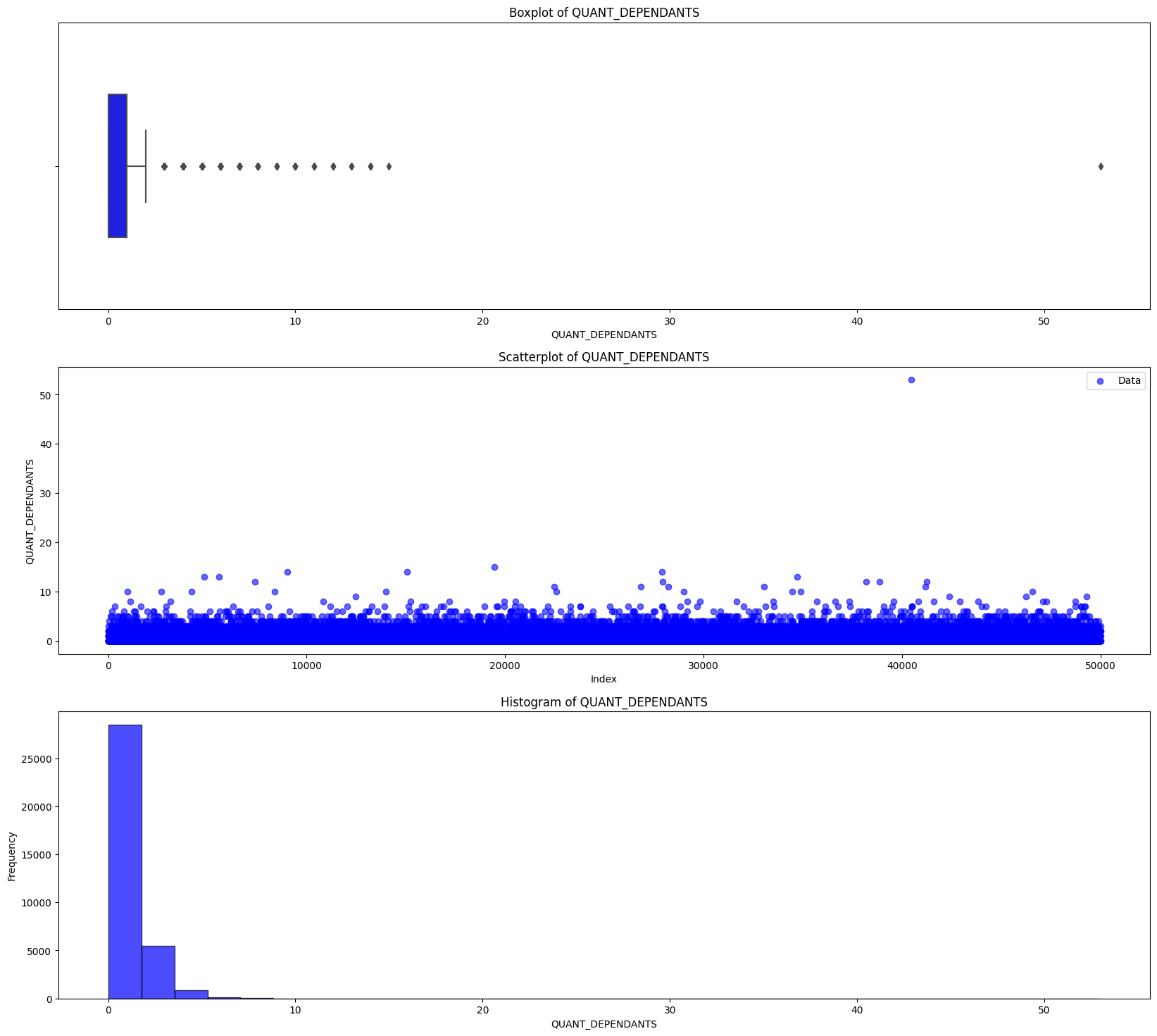
****For the variables MONTHS\_IN\_RESIDENCE, PERSONAL\_MONTHLY\_INCOME, and OTHER\_INCOME, we opted to retain all seemingly high but potentially reasonable values. These elevated figures are considered to represent important, real-life data scenarios, whether it pertains to long-term residence, high personal monthly income, or other incomes. This approach aims to preserve the realism of the dataset.

Figure X1: Outliers Visualisation in “QUANT\_DEPENDANTS”

For variable QUANT\_DEPENDANTS, we manually removed entries with values exceeding 50, as it's highly unlikely for an individual to have over 50 dependents. This decision is supported by the outlier visualisation (Figure X1), where data points significantly deviating from the normal range are clearly depicted.

Figure X2: Outliers Visualisation in “AGE”

From Figure X2, we can identify that the variable “AGE” is approximately normal-distributed, so we used z-SCORE to remove outliers.

## Feature Scaling

To apply the dataset to Logistic Regression, we used the standard scaler to standardize the numerical predictors.