**Multidimensional Data Analysis**

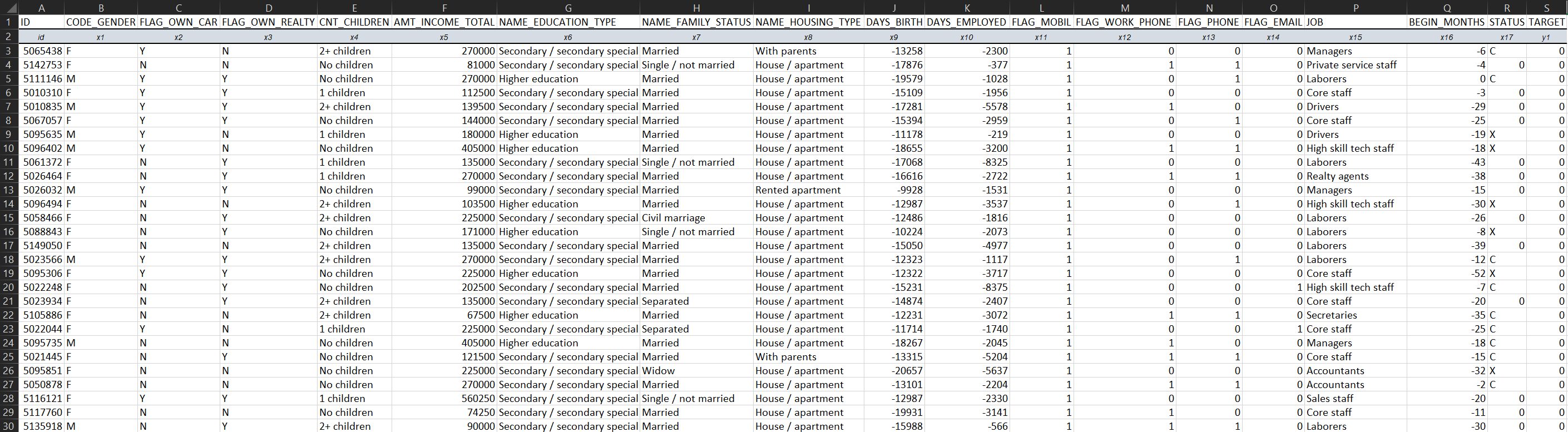
1. [Goal of Analysis](#GoalOfAnalysis)
2. [Description of Input Data](#DescriptionofInputData)
3. [Data Preparation](#DataPreparation)
4. [Initial Data Analysis](#InitialDataAnalysis)
5. [Data Visualization](#DataVisualization)
6. [Division of the set of objects into two groups](#DivisionOfTheSetOfObjectsIntoTwo)
7. [Artificial Neural Network (ANN) Model](#ANN_Model)

**1. Goal of Analysis**

The goal of this project is to analyze the provided credit card approval data to identify patterns and relationships between the applicant's information and the likelihood of their credit card application being approved. This analysis will aim to develop predictive models that assess the probability of approval based on various factors.

**2.** **Description of Input Data**

The input data consists of a file containing information about credit card applicants. The dataset includes variables such as applicant's age, income, employment length, education, marital status, home ownership, credit score, and loan amount. The target variable is the binary response of whether the applicant's credit card application was approved or not.



***X’s attributes:***

**ID:** Client Number

**CODE\_GENDER:** Gender

**FLAG\_OWN\_CAR:** Is there a car

**FLAG\_OWN\_REALTY:** Is there a property

**CNT\_CHILDREN:** Number of Children

**AMT\_INCOME\_TOTAL:** Annual Income

**NAME\_EDUCATION\_TYPE:** Education Level

**NAME\_FAMILY\_STATUS:** Marital Status

**NAME\_HOUSING\_TYPE:** Way of Living

**DAYS\_BIRTH:** Age in days

**DAYS\_EMPLOYED:** Duration of work in days

**FLAG\_MOBIL:** Is there a mobile phone

**FLAG\_WORK\_PHONE:** Is there a work phone

**FLAG\_PHONE:** Is there a phone

**FLAG\_EMAIL:** Is there an email

**JOB:** Job

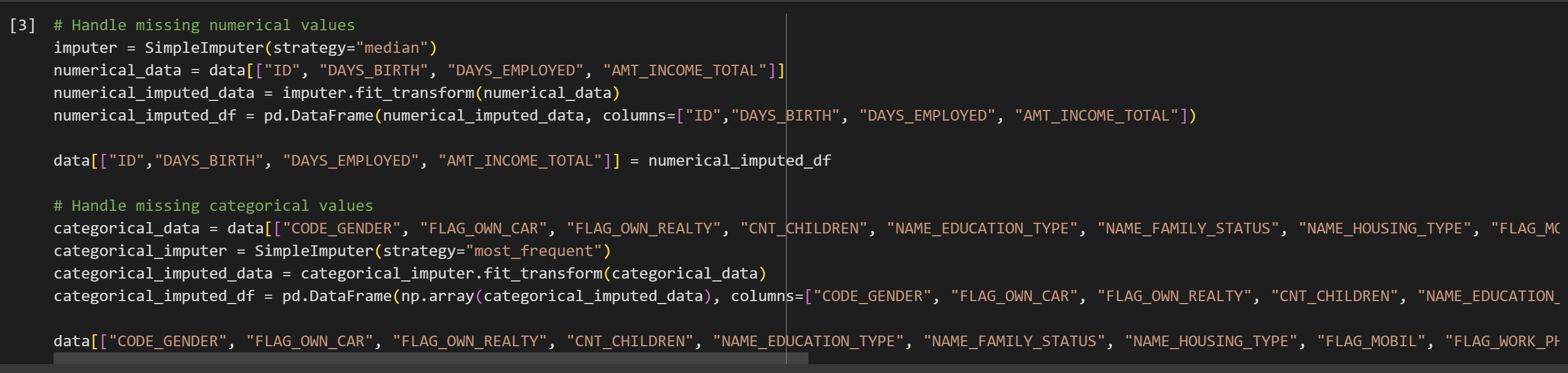
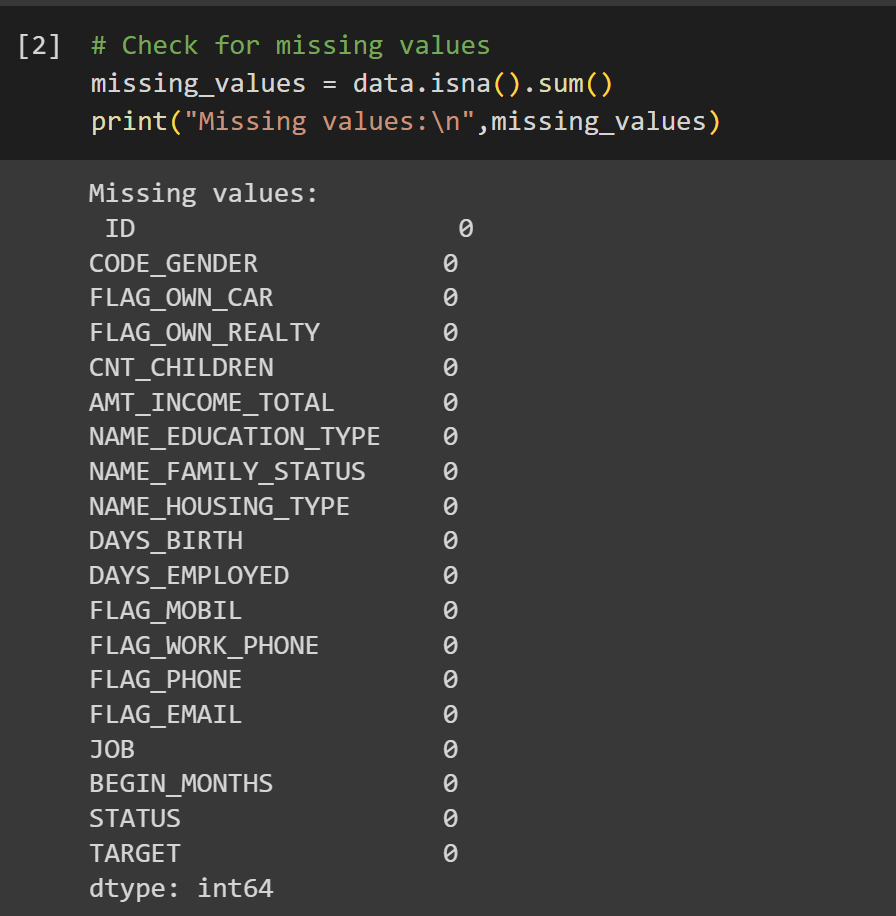
**BEGIN\_MONTHS:** Record month (The month of the extracted data is the starting point, backwards, 0 is the current month, -1 is the previous month, and so on)

**STATUS:** Status (0: 1-29 days past due 1: 30-59 days past due 2: 60-89 days overdue 3: 90-119 days overdue 4: 120-149 days overdue 5: Overdue or bad debts, write-offs for more than 150 days C: paid off that month X: No loan for the month)

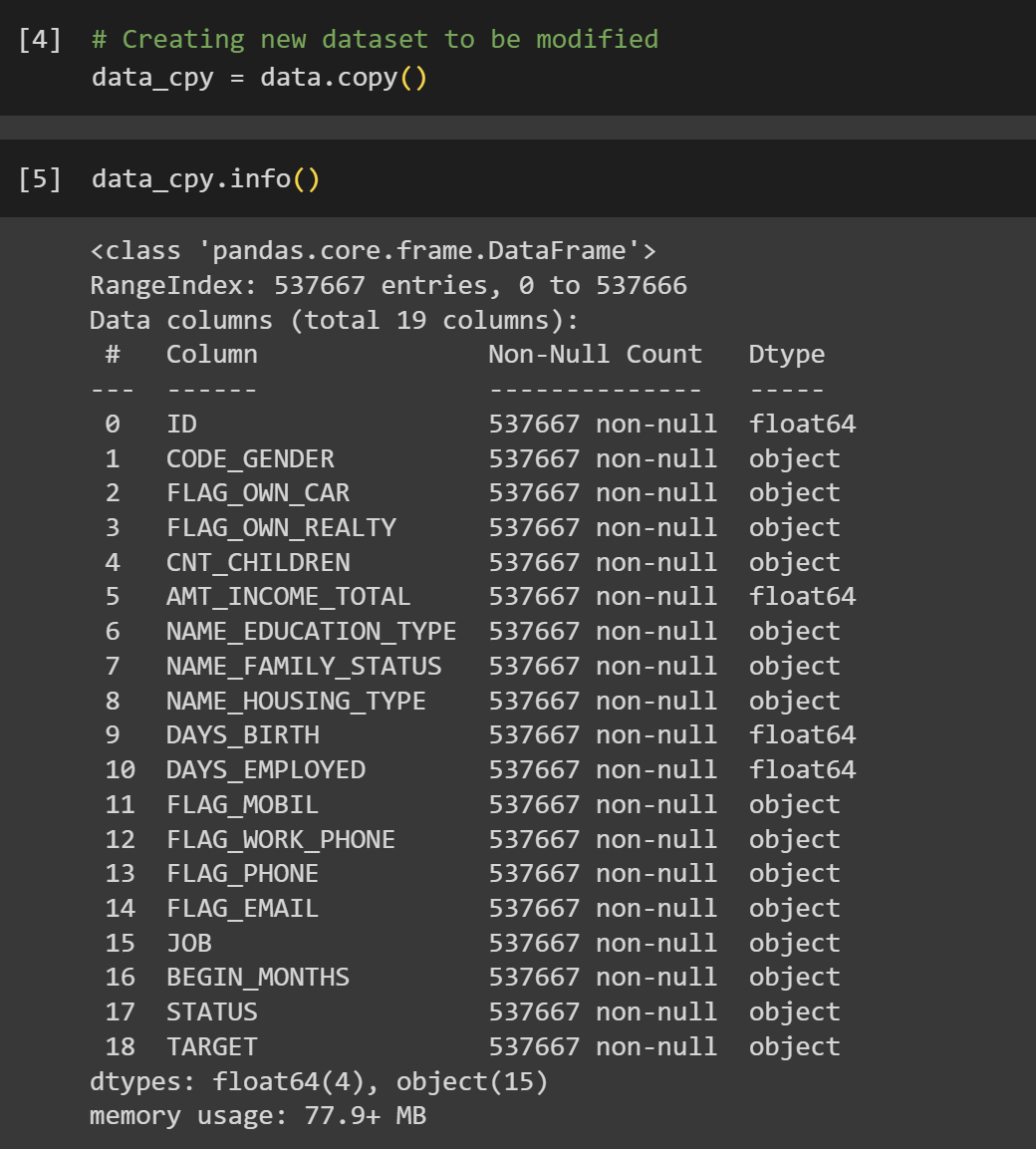
***Y’s attributes:***

**TARGET:** Target (Risk user are marked as '1', else are '0')

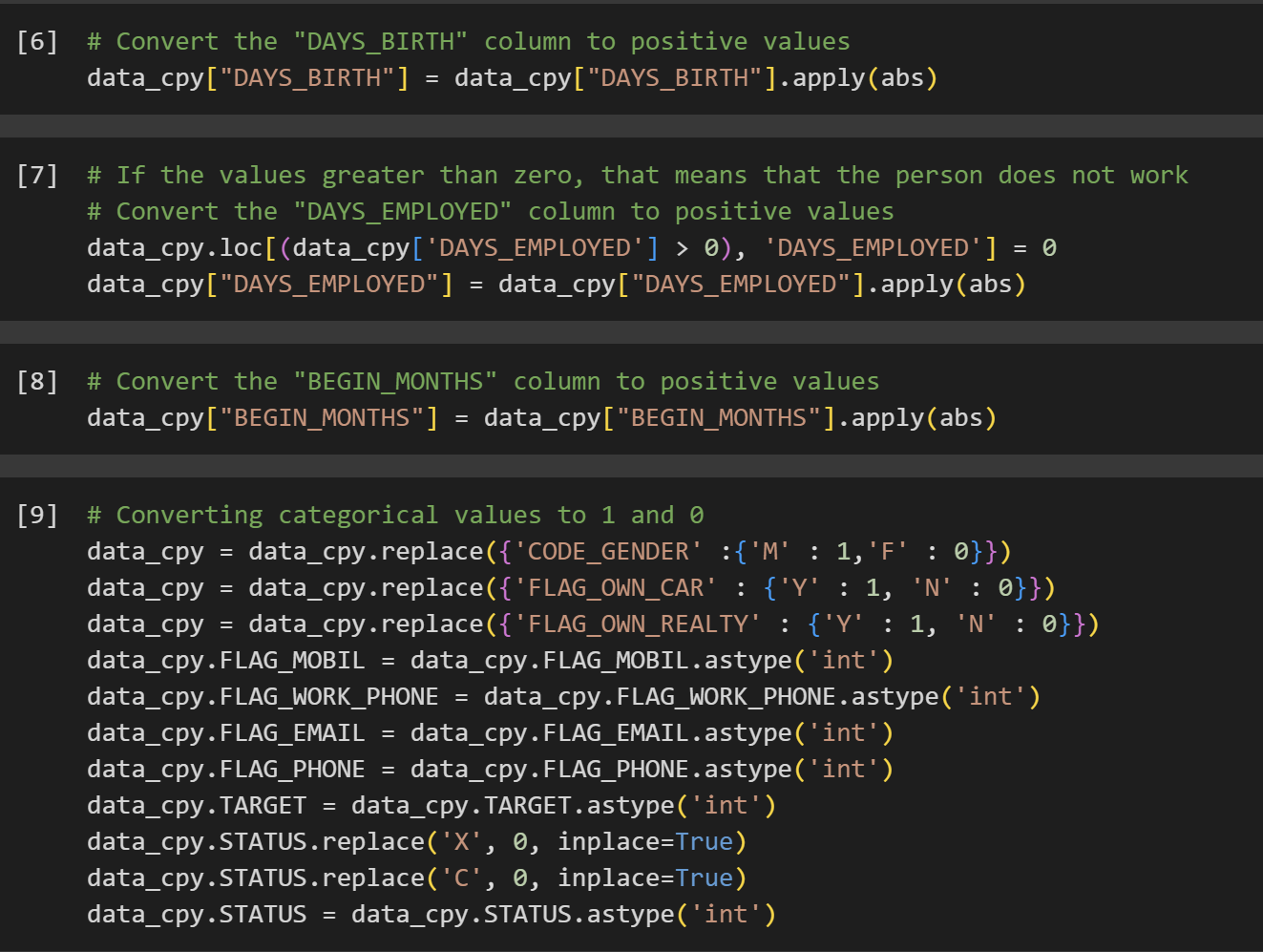
**3.** **Data Preparation**



This section is for preparing the credit card approval data for further analysis. Initially we identifying missing values and using appropriate techniques to impute them (There is no missing data but I still implemented the code for Handling missing data just in case). For numerical variables, the median is used, while for categorical variables, the most frequent category is imputed.

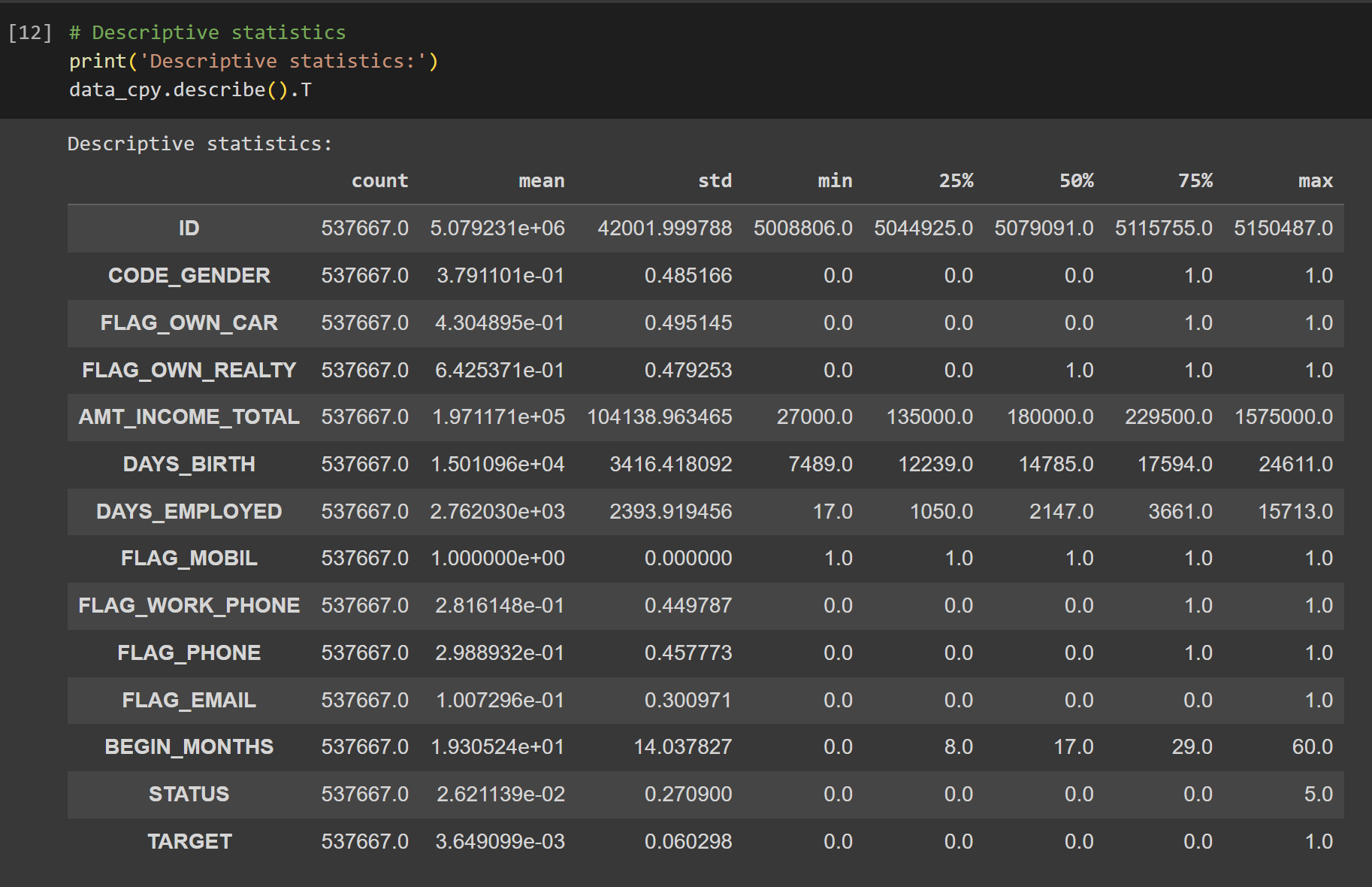


At this step, I created a new dataset to be used for further analysis so the initial one to remains untouched.



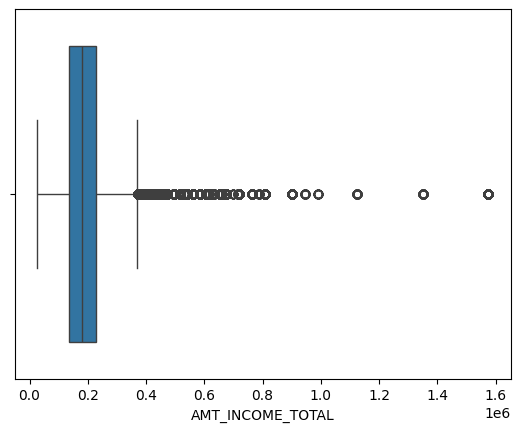
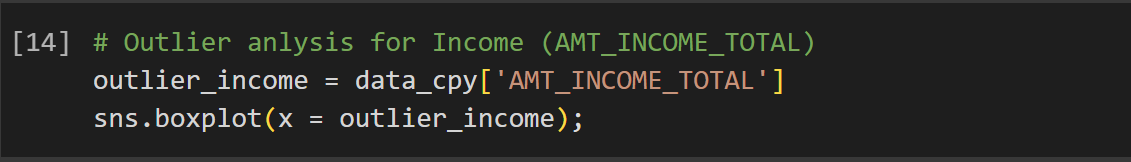
This code focuses on further preprocessing the data after handling missing values. It aims to transform categorical variables into numerical representations that can be used for analysis.

**4:** **Initial Data Analysis**



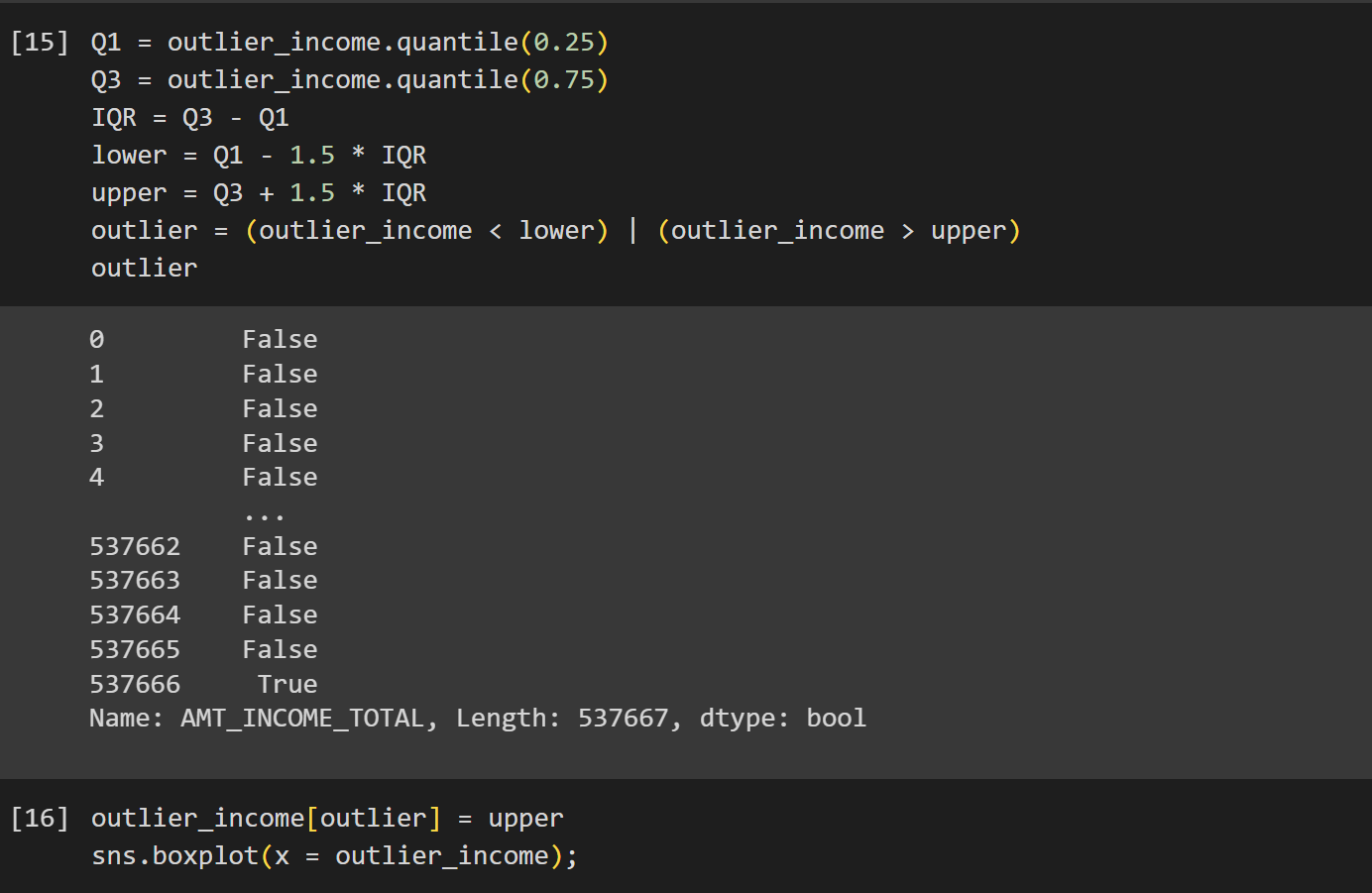
This part of the program is displaying the descriptive statistics for the numerical variables in the dataset. This includes the count, mean, standard deviation, minimum, 25th percentile, 50th percentile, 75th percentile, and maximum for each numerical variable.

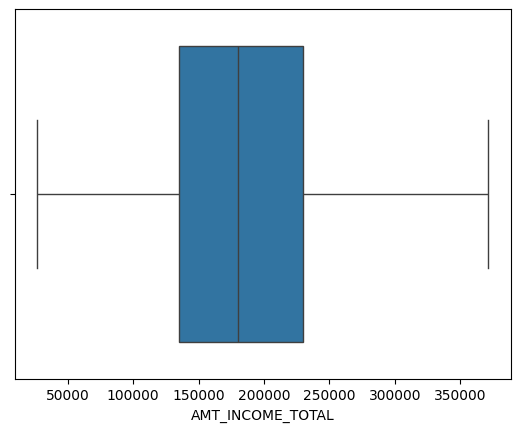
At the beginning step of the data analysis, we filter out the data by detecting the outliers. Outliers can distort the distribution of the data and can affect the accuracy of the machine learning models. By removing the outliers, we have more suitable data for analysis and modeling.



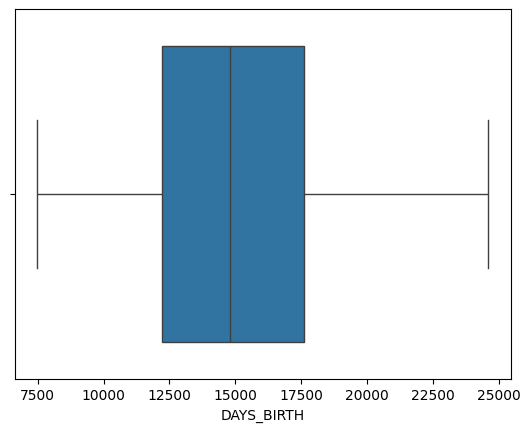
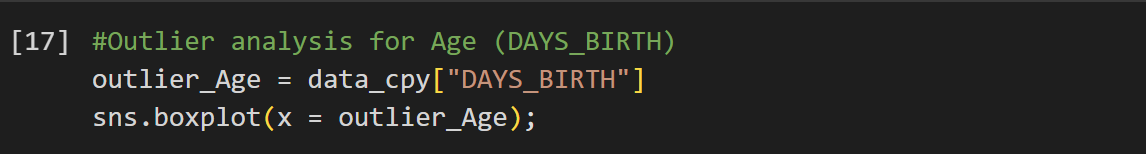
This part of the code displays a box plot showing that there are outliers in the AMT\_INCOME\_TOTAL variable, which is the Annual income. The box plot shows the median, 25th percentile, 75th percentile, and minimum and maximum values for the variable. The outliers are the points that fall outside the whiskers, which extend from the 25th percentile to the 75th percentile plus 1.5 times the interquartile range (IQR). In this case, the outliers are the points that are below 0.2 and above 1.6. This suggests that there may be some data errors or anomalies in the AMT\_INCOME\_TOTAL variable.

Same principle applies also to the ones from below.

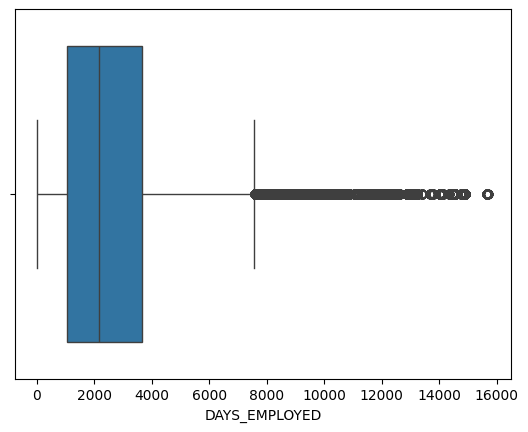
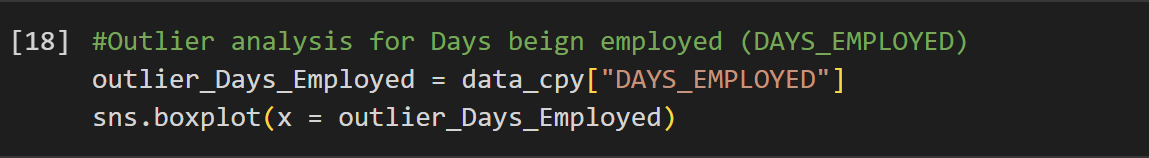




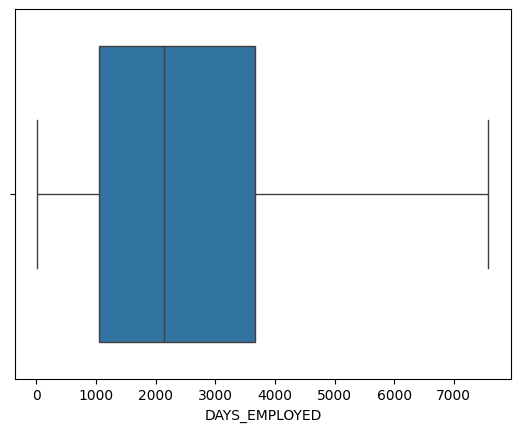
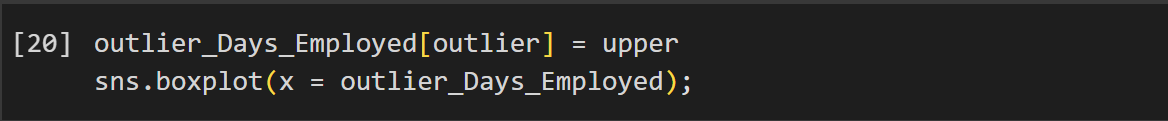
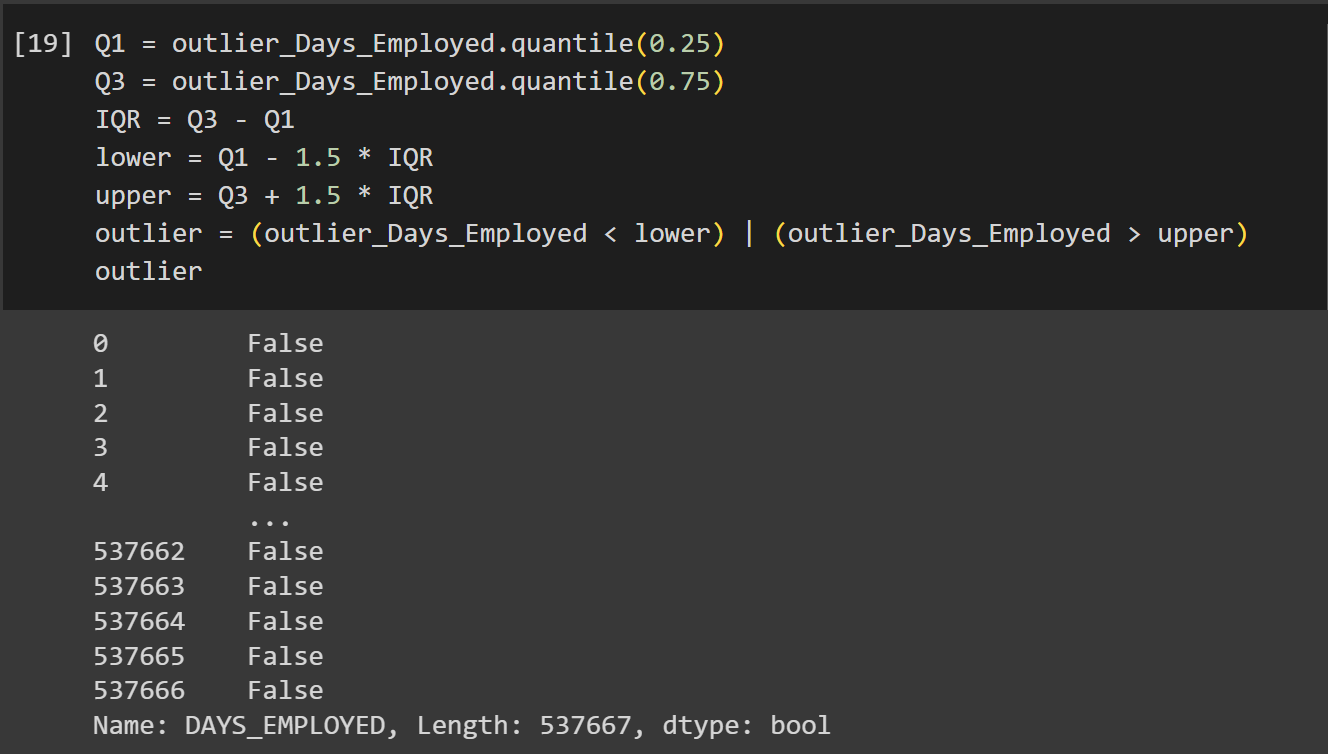
The code snippet identifies and removes outliers from the AMT\_INCOME\_TOTAL variable using the Interquartile Range (IQR) method. After performing the removal of the outliers, the program shows the updated plot box.



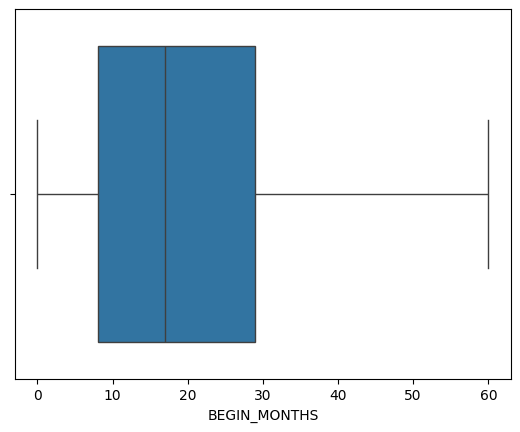
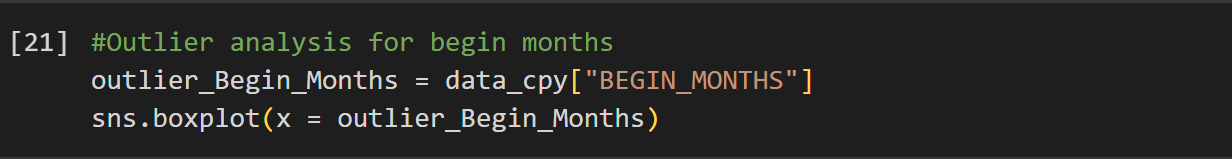
There are no outliers for the DAYS\_BIRTH variable.



The box plot with the outliers for the variable DAYS\_EMPLOYED.



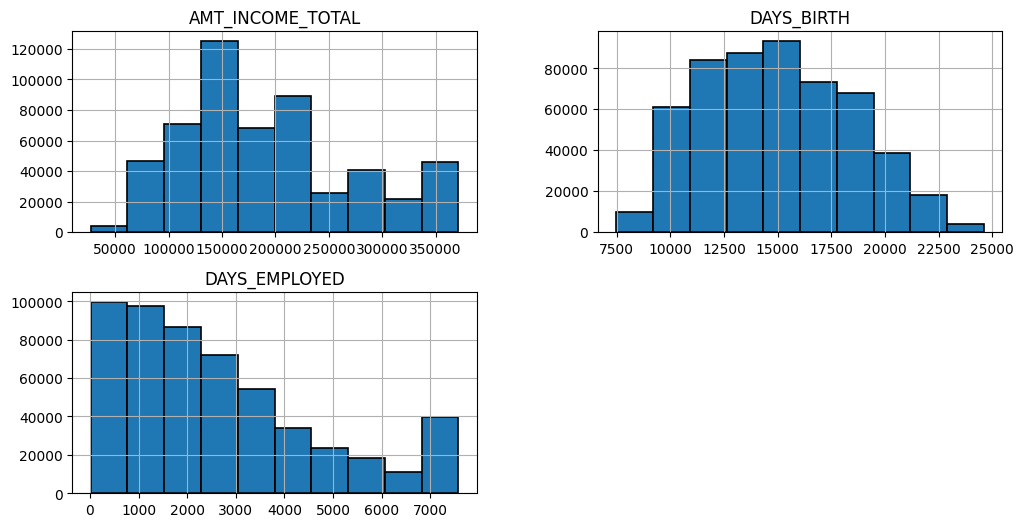
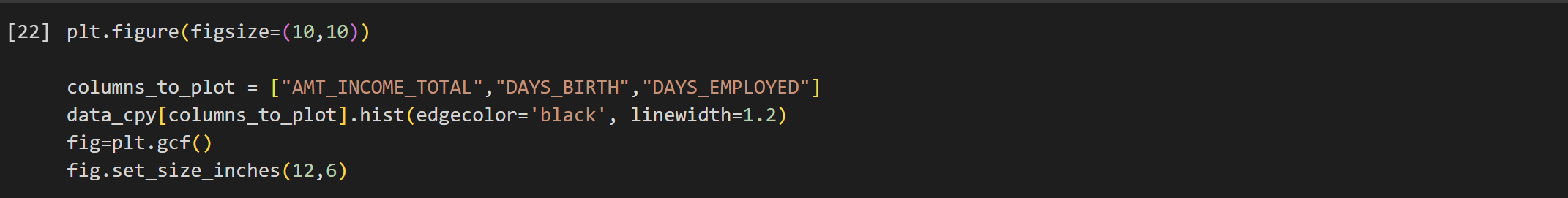
The box plot with the outliers removed for the variable DAYS\_EMPLOYED.

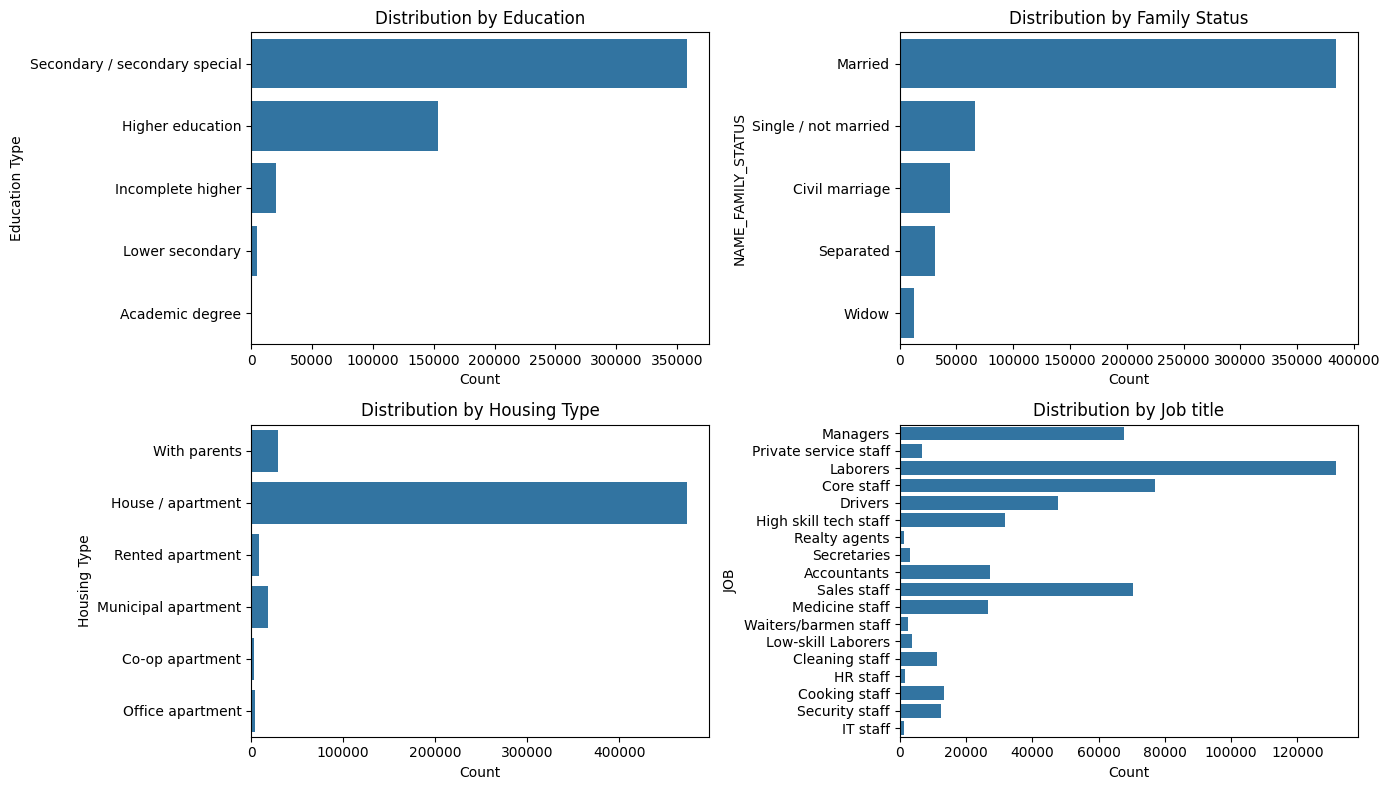
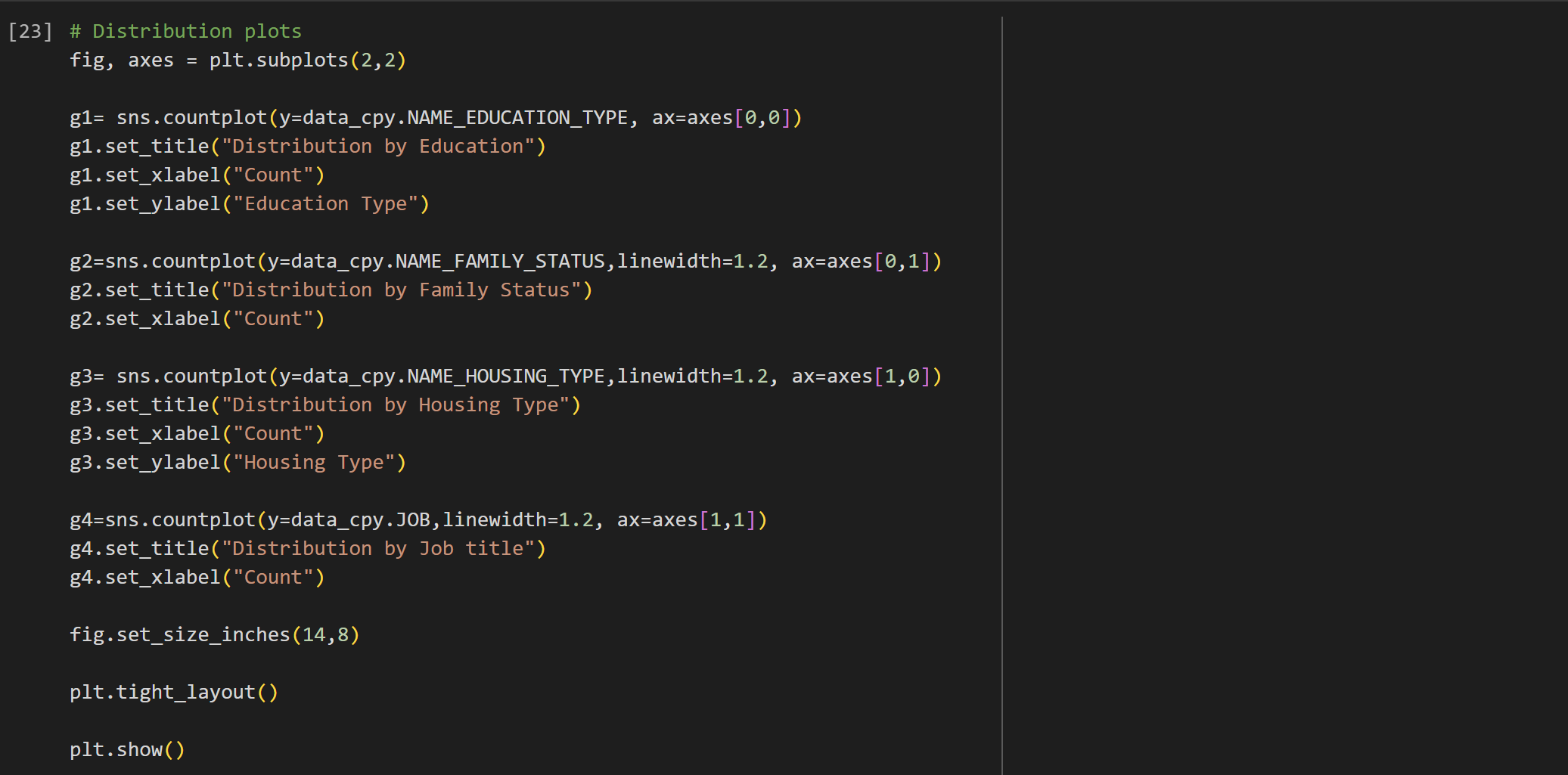


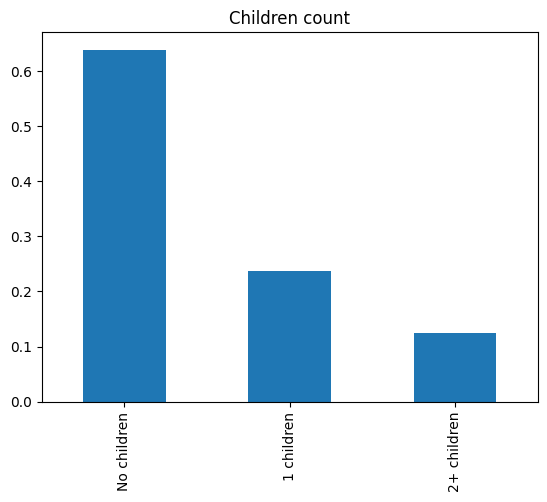
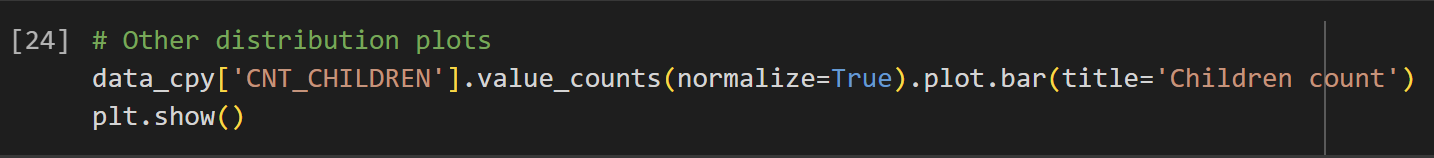
There are no outliers for the BEGIN\_MONTHS variable.

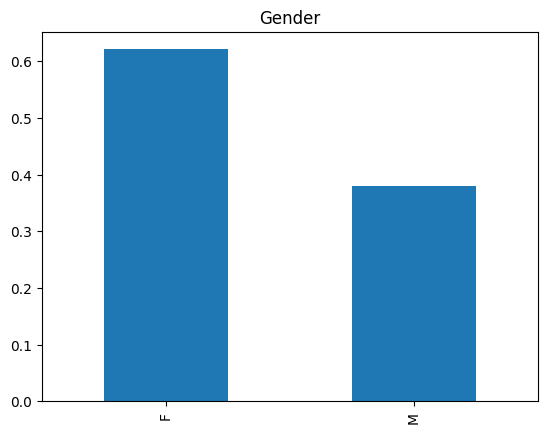
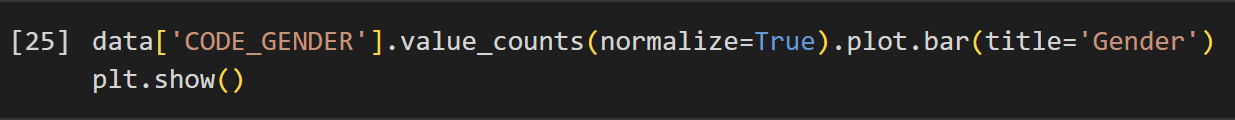
**5:** **Data Visualization**

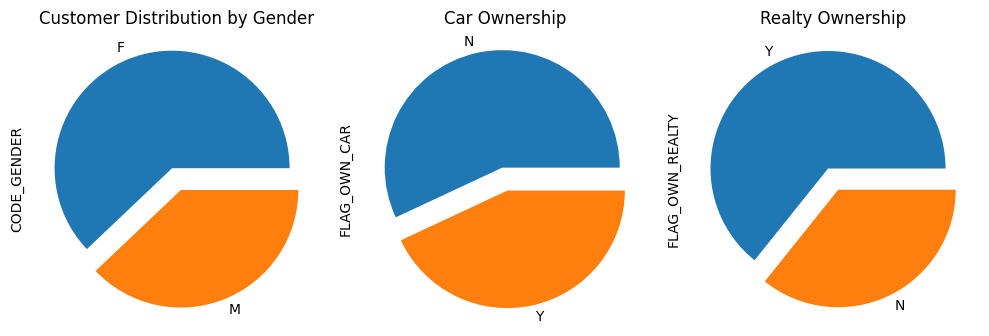
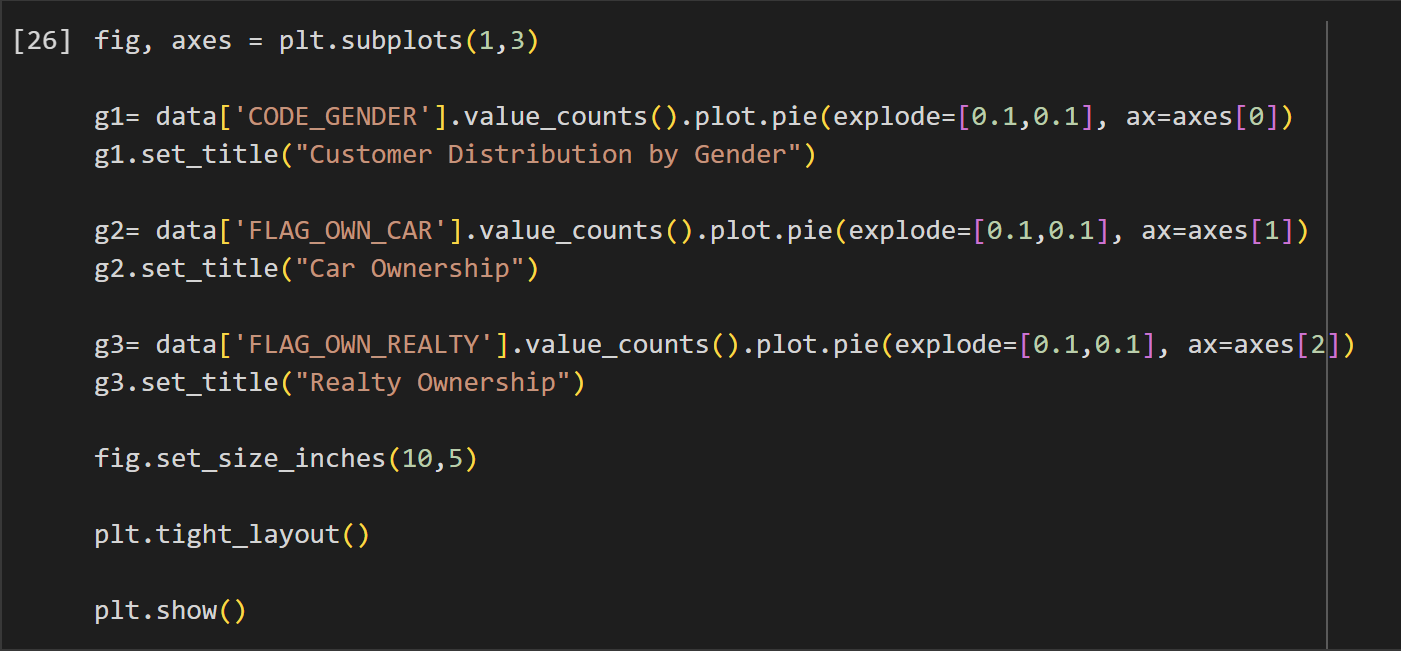
Under this section I start displaying the data visualization of the variables in the dataset.

****

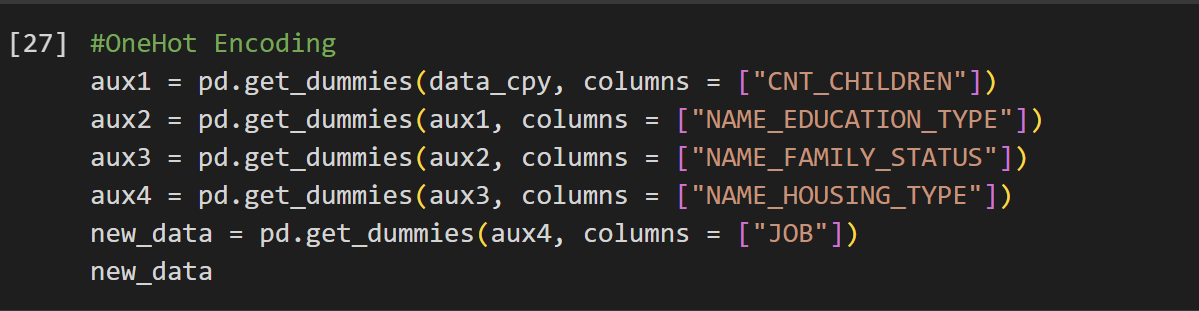




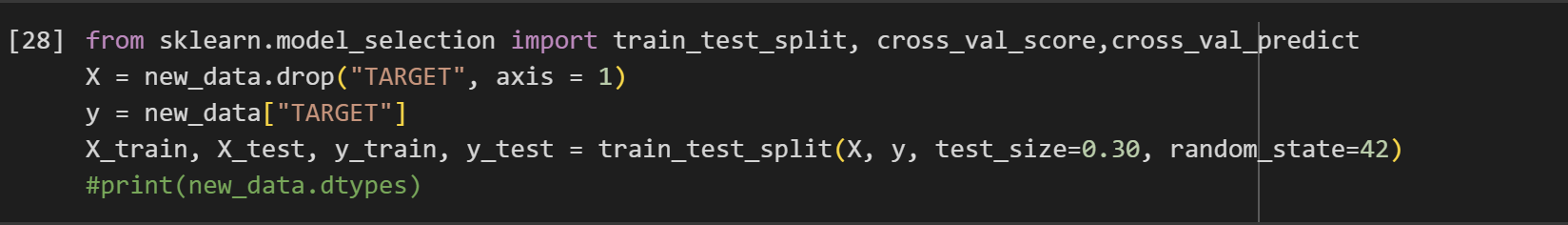




**6:** **Division of the set of objects into two groups**



In this part of the code, we transform categorical variables into numerical representations. The code utilizes the One-Hot Encoding technique, which creates a binary variable for each unique value in a categorical column.



Here we perform train-test split on the dataset to evaluate the performance of the machine learning model. The code randomly splits the dataset into two parts: training and testing sets.

**7:** **Artificial Neural Network (ANN) Model**

*Theoretical basis of the model and an analysis performed on sample data:*

*ANNs are powerful learning models that can effectively capture complex relationships in data. They have demonstrated remarkable performance in various classification tasks, particularly those with nonlinear relationships. In the code we have the implementation of an ANN model using the MLPClassifier class for training a binary classifier.*

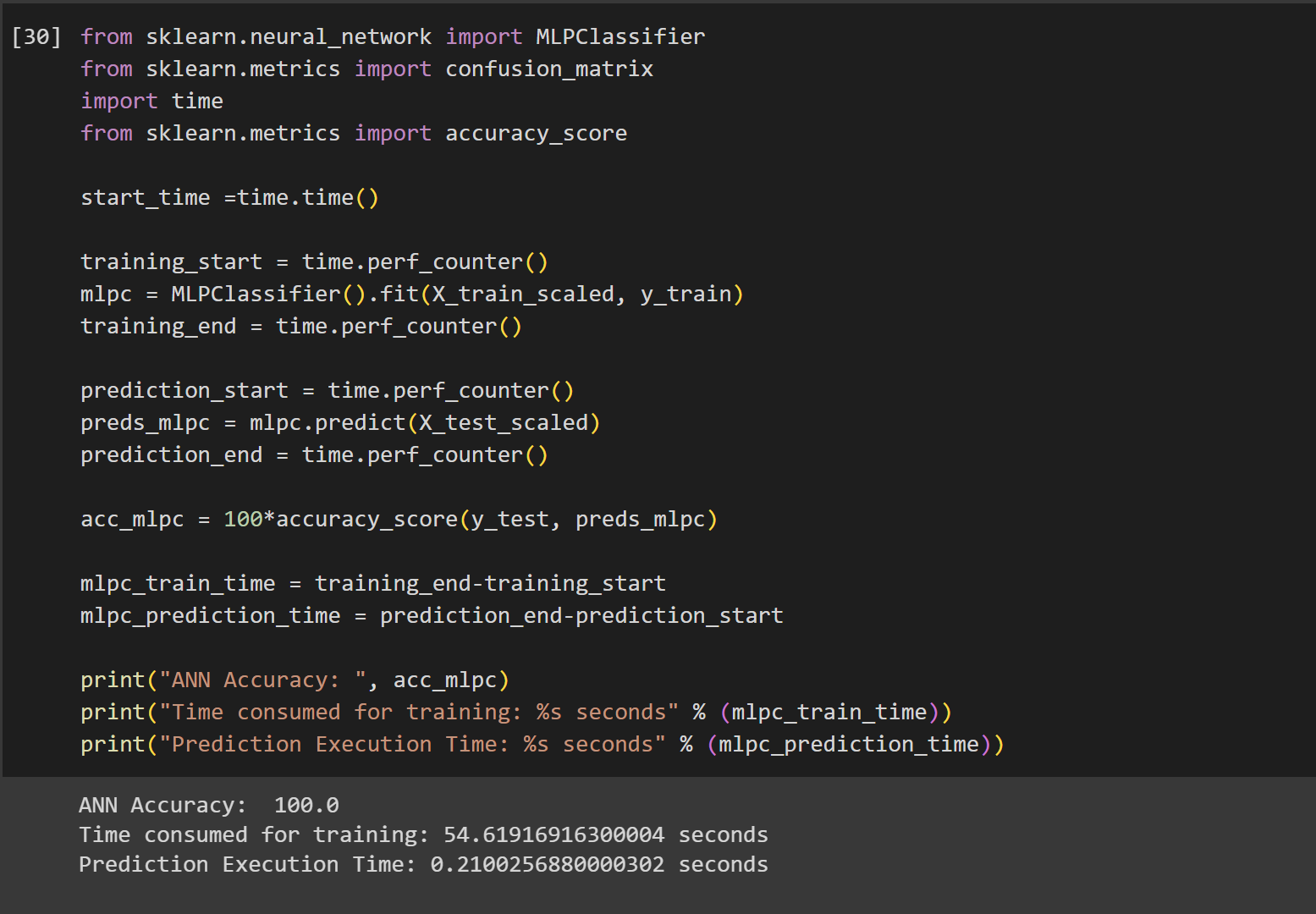
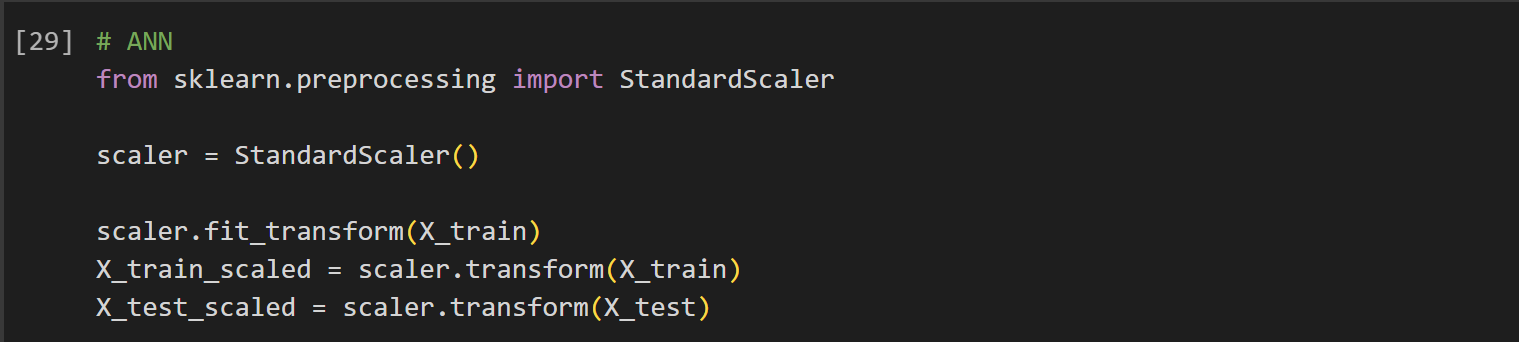
*Summary of the code's ANN-specific components:*

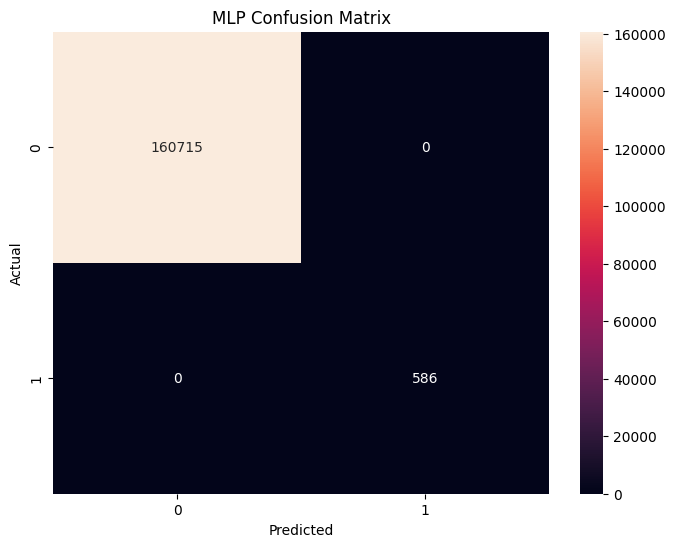
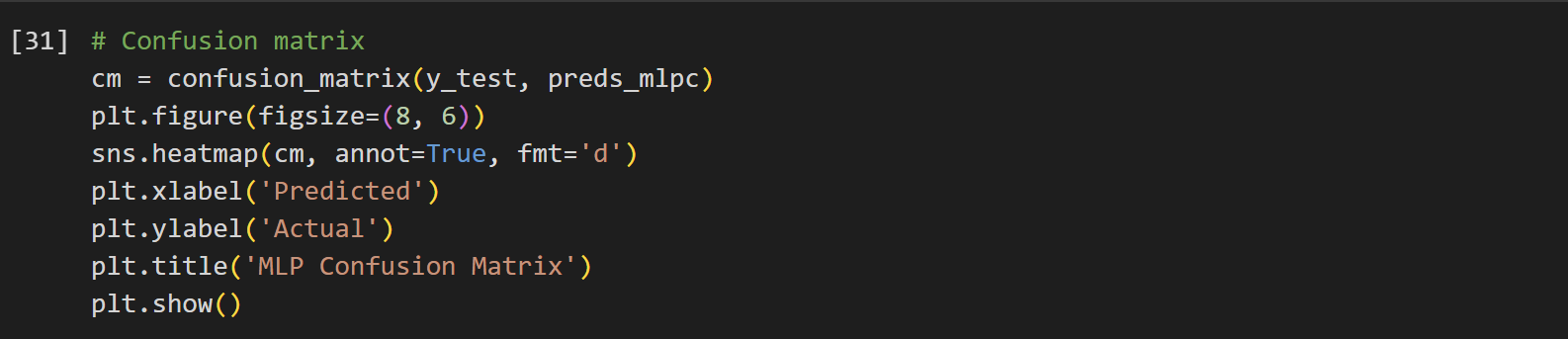
*-* *Import MLPClassifier*

*- Train the ANN*

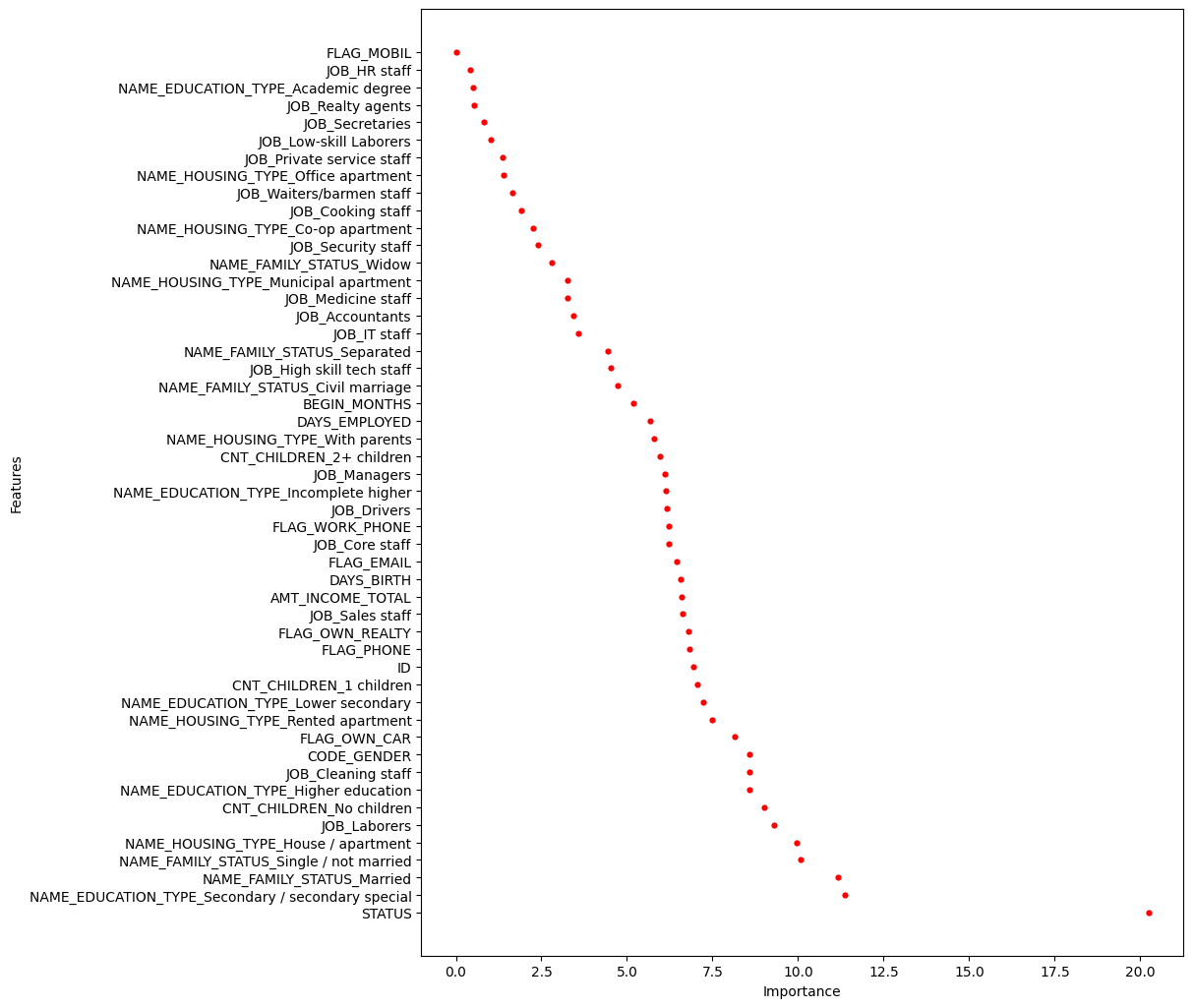
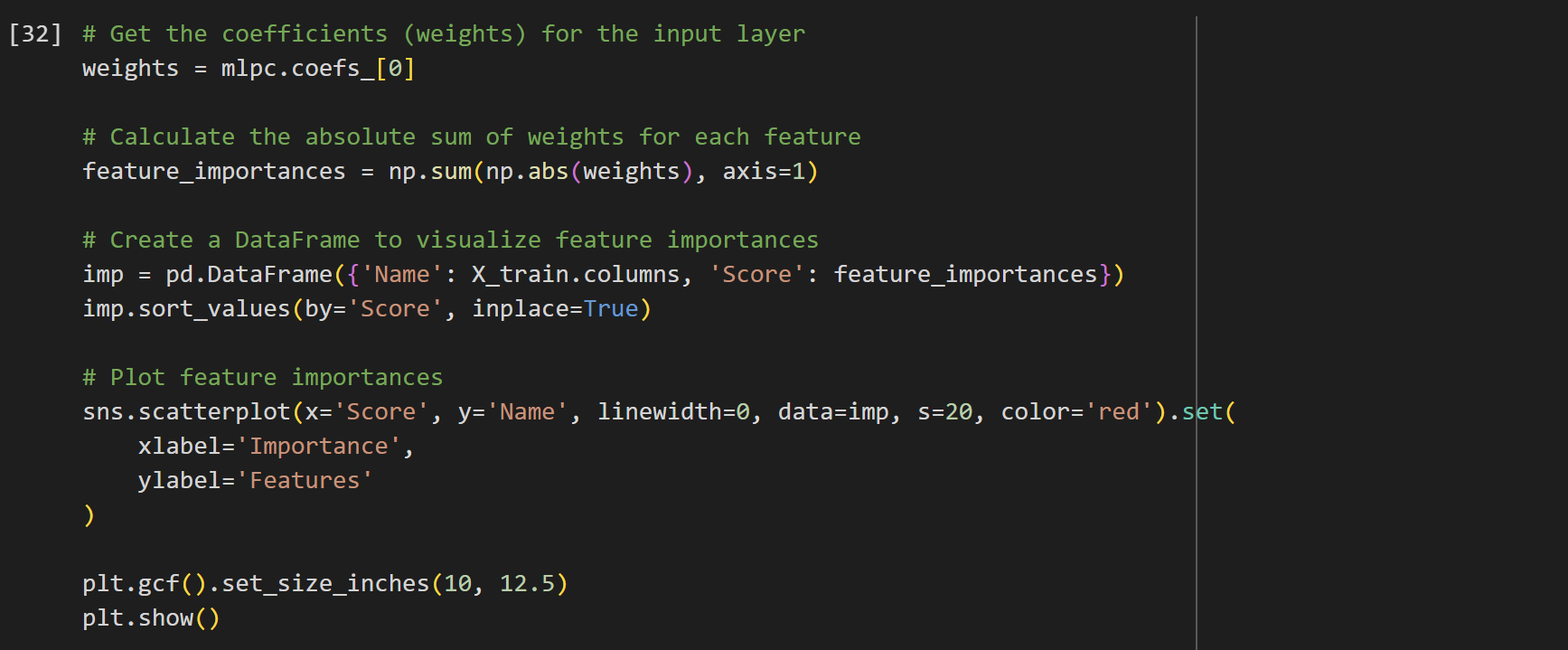
*- Make predictions*

*- Evaluate accuracy*

This code here demonstrates the implementation of a multilayer perceptron (MLP) classifier, a type of artificial neural network (ANN), for training a binary classifier. The output presents the accuracy, the time consumed for training and the predicted execution time.



The provided graph illustrates the performance of an artificial neural network (ANN) model in classifying data. It visualizes the model's ability to accurately predict the target variable based on the input features. The ANN's performance is evaluated using the accuracy metric, which represents the percentage of correctly classified instances.



This part shows the feature importance for the artificial neural network (ANN) model trained to classify data. It visualizes how each input feature contributes to the model's predictions.