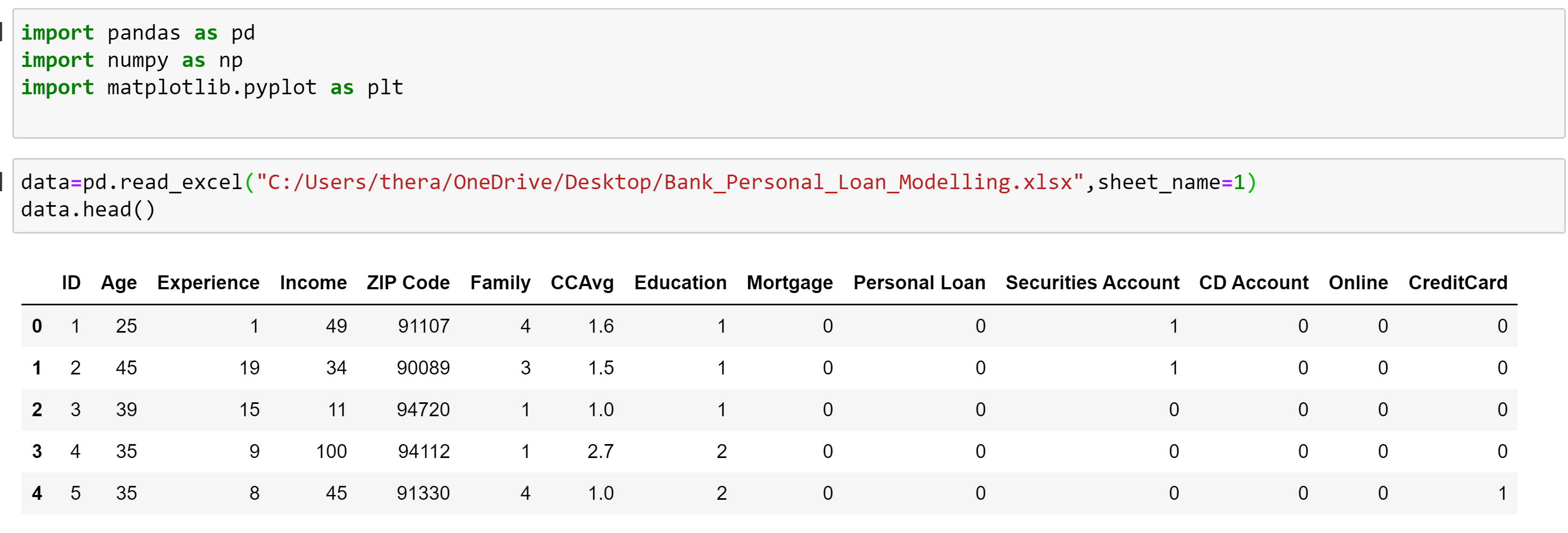
**Bank Loan modelling project**

**Step 1:** Loading the dataset using pandas and using head method to check whether data is properly loaded or not.



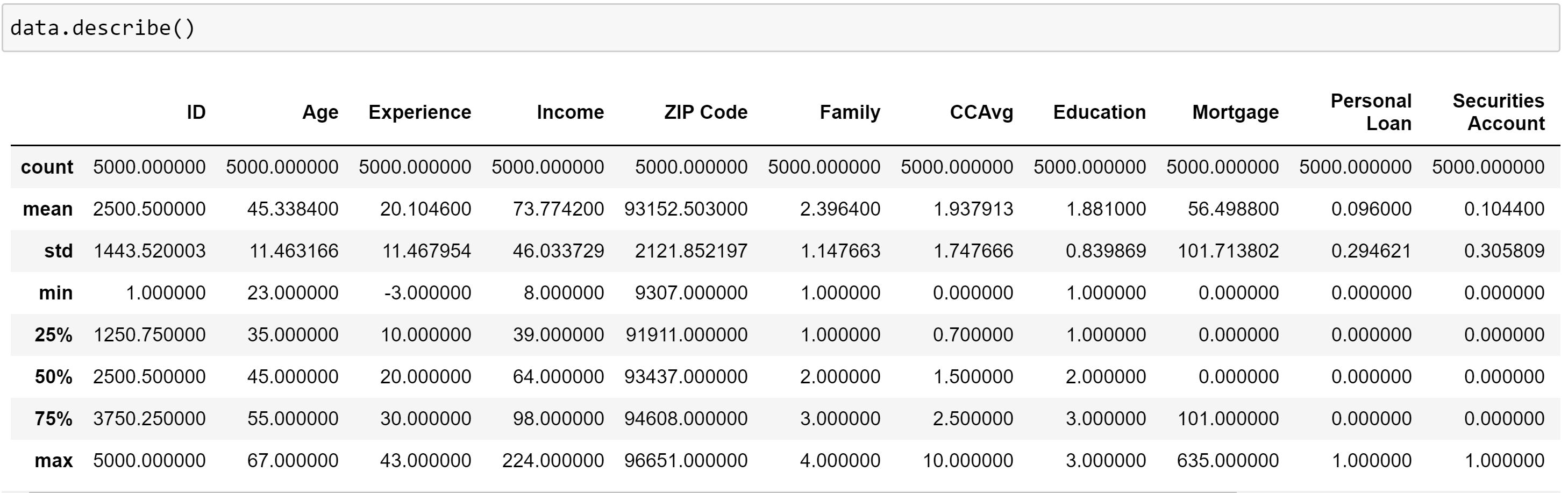
**Step 2:** Data Cleaning

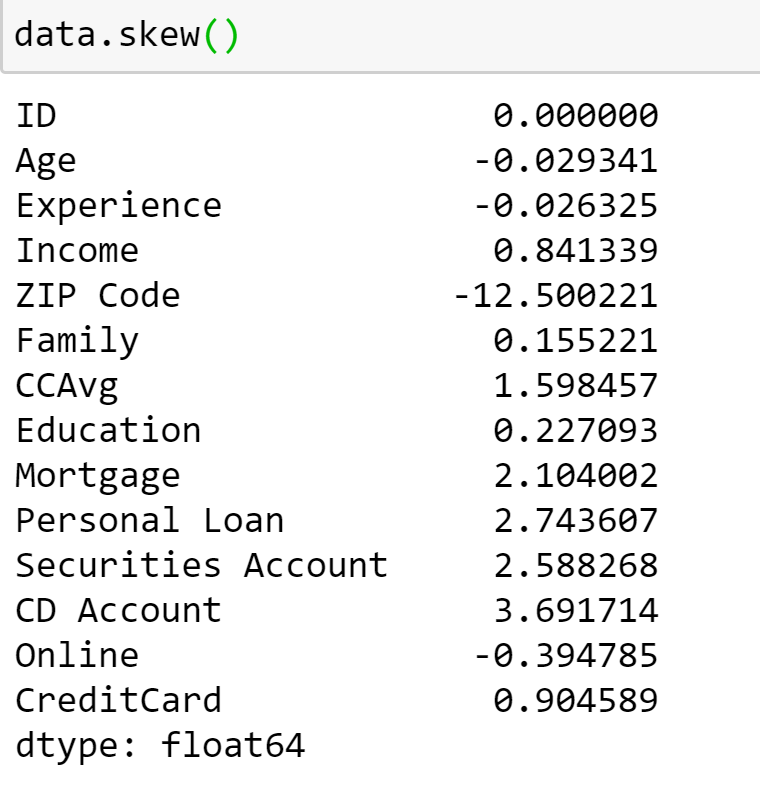
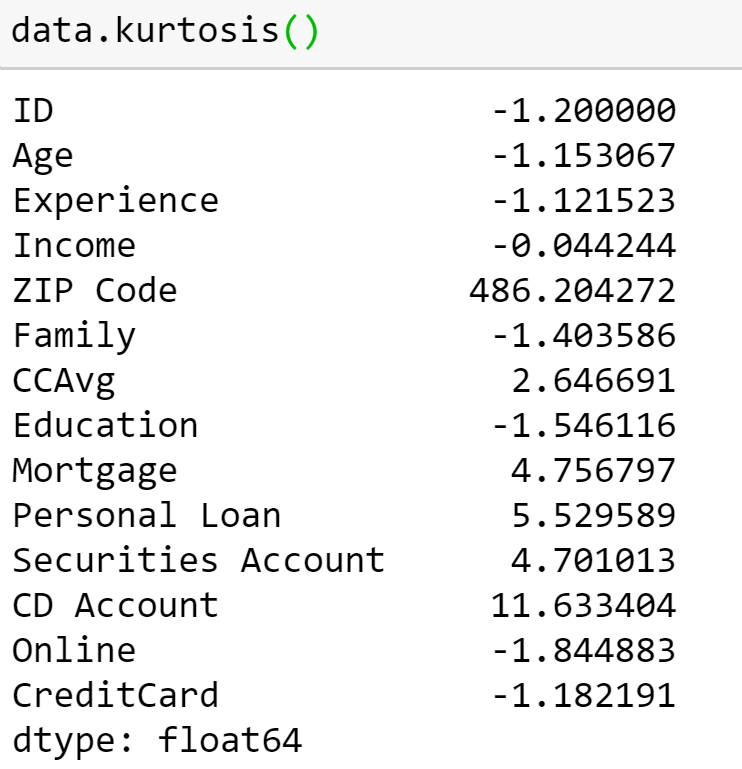
Checking the dataset for any missing values and duplicated data is dropped from the dataset. So , the bank modeling dataset have no missing values.



**Descriptive Statistics:**

**Step 3:** Describe method displays the summary statistics of all the variables such that it is easy to know how each variable is distributed. From the below snip , we can say that income mean is greater than median so it has positive skew , age variable is normally distributed because mean and median are almost equal. Even education has approximate normal distribution.



From the above snips, we can say that variables age,experience,zipcode and online are negatively skewed and remaining variables have positive skew. Out of all the variables, zip code is more negatively skewed and cd account has more positive skew.

Coming to the kurtosis, zip code is more leptokurtic and most of the variables are platykurtic.

**Step 4: Correlation analysis:**

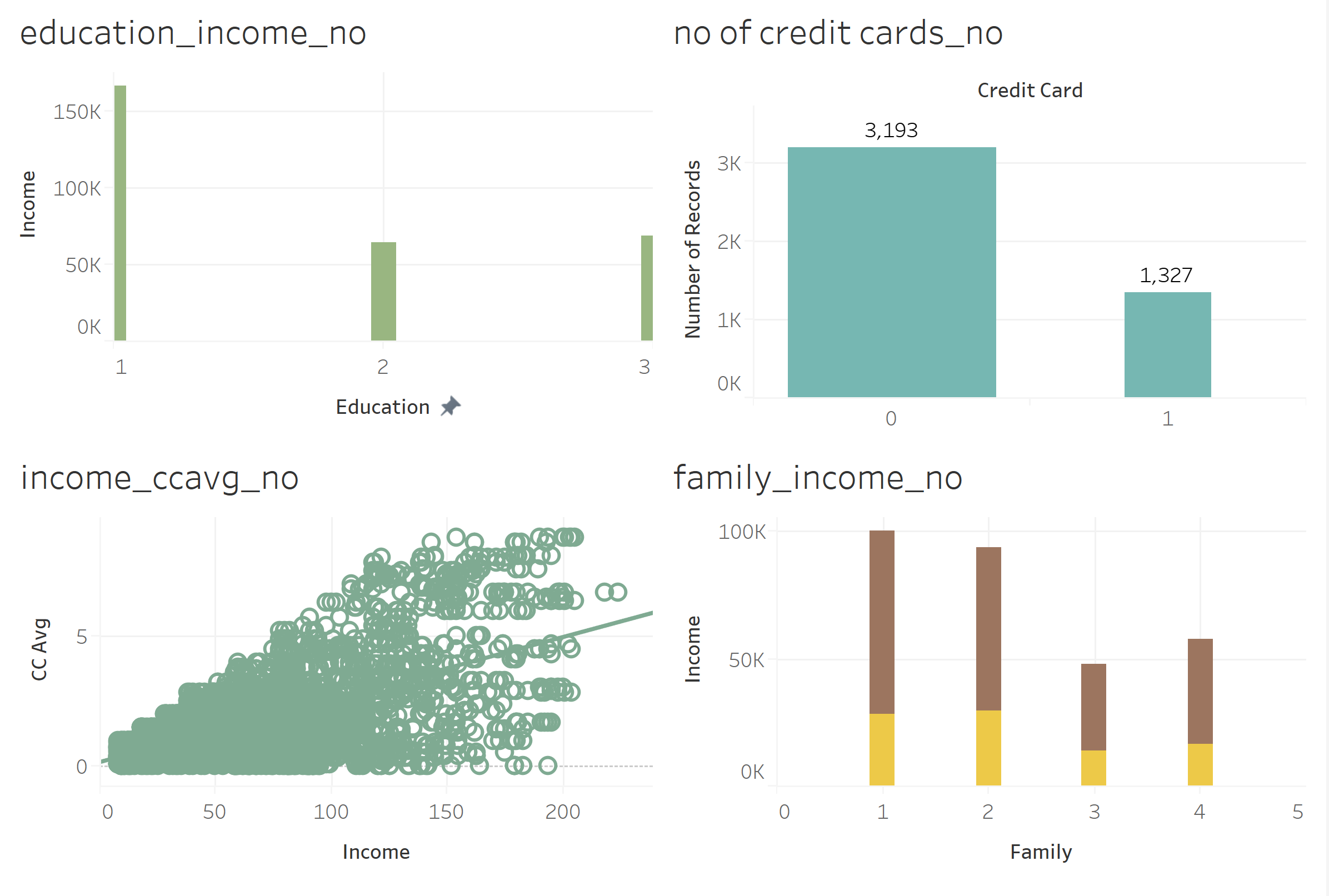
Data.corr() method displays the correlation matrix that is how one variable is correlated with the other .From the below result we can see that all the variables have weak correlation with the personal loan except income , ccavg and followed by cd account.



**Step 5:**

**Tableau Visualizations**

**Plot 1:** Below is the dashboard about the people who did not get the personal loan from the bank.

**Conclusions:** About the people who did not get the personal loan from bank:

1.From the first plot, we can say that undergrad people are earning more when compared to masters and advanced level.

2.Credit cards are not issued to most of the people.

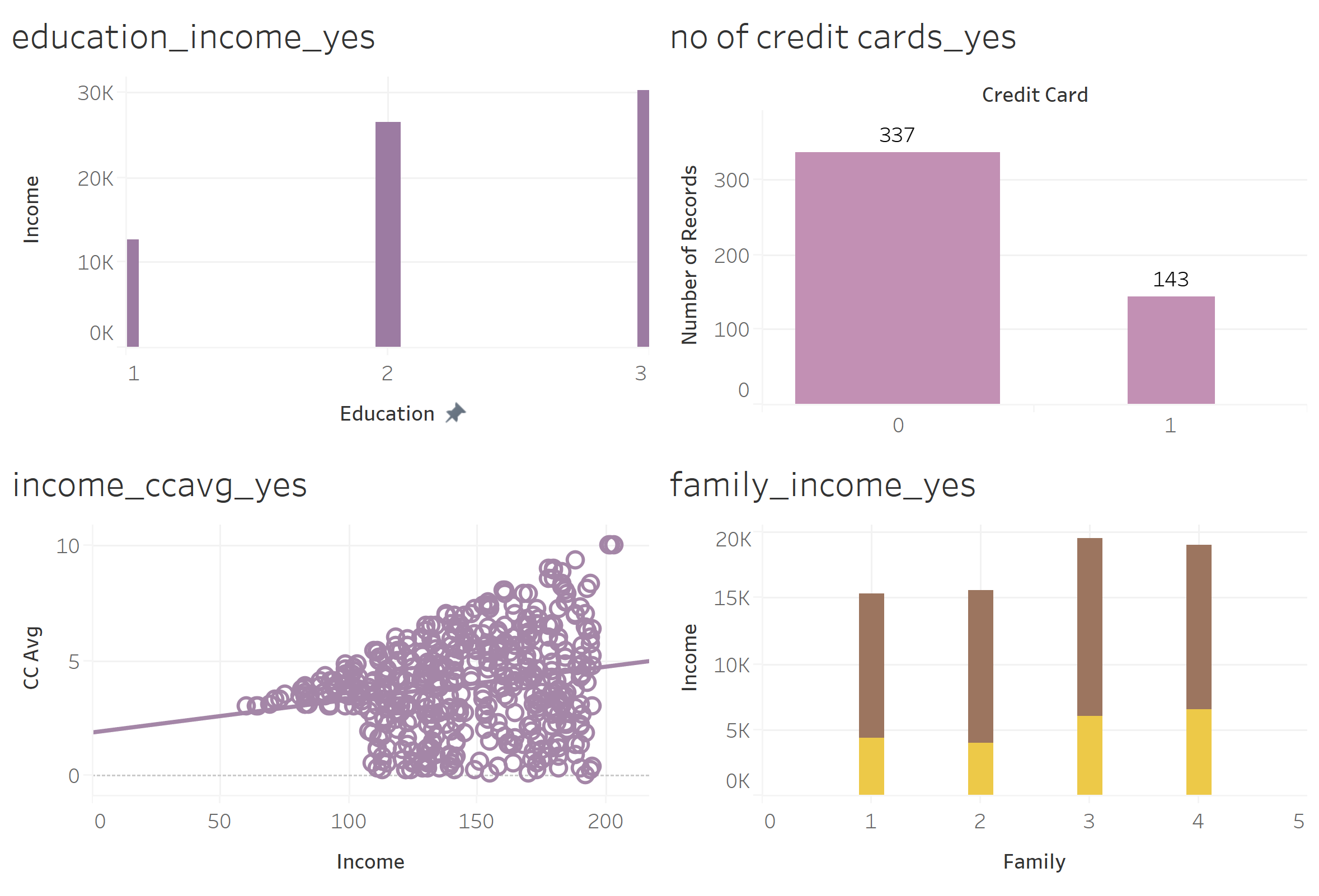
3.Graph between income and ccavg shows the linear trend with Rsquare value of 0.38 and p value is less than 0.05 that means it is significant. So , we can say that as ccavg increases , income is also increasing.

|  |  |
| --- | --- |
| **P-value:** | < 0.0001 |
| **Equation:** | CC Avg = 0.0240342\*Income + 0.137018 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coefficients** | | | |  |
| **Term** | **Value** | **StdErr** | **t-value** | **p-value** |
| Income | 0.0240342 | 0.00045 | 53.4112 | < 0.0001 |
| intercept | 0.137018 | 0.0349533 | 3.92004 | < 0.0001 |
|  |  |  |  |  |

4.Family with one person is earning more when compared to the other(comparing length of the bar) and more no of credit cards are issued to the family of size 2 ( indicated by yellow bar).Family of size 3 has low income and credit cards issued are less when compared to all other family sizes.

**Plot 2:** Below is the dashboard about the people who got the personal loan from the bank.

**Conclusions:**

1.Advanced level/professionals earn more when compared to undergrad and master’s people.

2. Number of credit cards issued and not issued are very less when compared to the people who did not get the loan.

3.Graph between income and ccavg shows the linear trend with R squared value of 0.04 and p value is less than 0.05 that means variable is significant. As R square is very less for the model, linear regression is not suitable for this type of data.

|  |  |
| --- | --- |
| **P-value:** | < 0.0001 |
| **Equation:** | CC Avg = 0.0142781\*Income + 1.83866 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coefficients** | | | |  |
| **Term** | **Value** | **StdErr** | **t-value** | **p-value** |
| Income | 0.0142781 | 0.0029667 | 4.81275 | < 0.0001 |
| intercept | 1.83866 | 0.439504 | 4.18348 | < 0.0001 |

4.Family with size 3 and 4 earn more when compared to the other families and a greater number of credit cards are issued to the family size 4.

**Step 6: Inferential Statistics:**

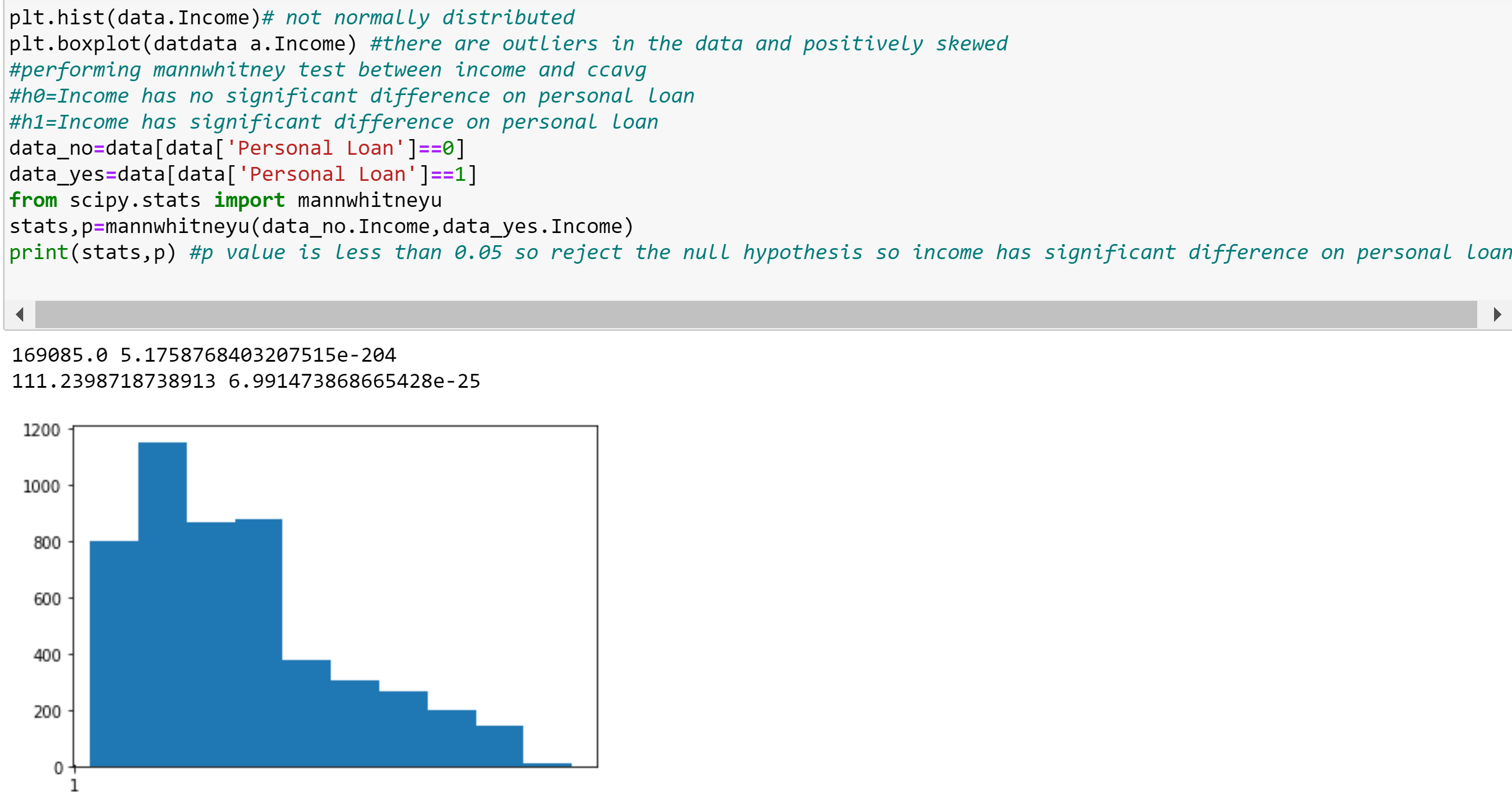
**1.**To decide the test, we need to check the distribution of income variable.

From the below plot you can see that, income is positive skewed so we need to perform non-parametric test.

Performing mannwhitney test between income and personal loan

h0=Income has no significant difference on personal loan

h1=Income has significant difference on personal loan



P value is less than 0.05 so reject the null hypothesis that means income has significant affect on personal loan.

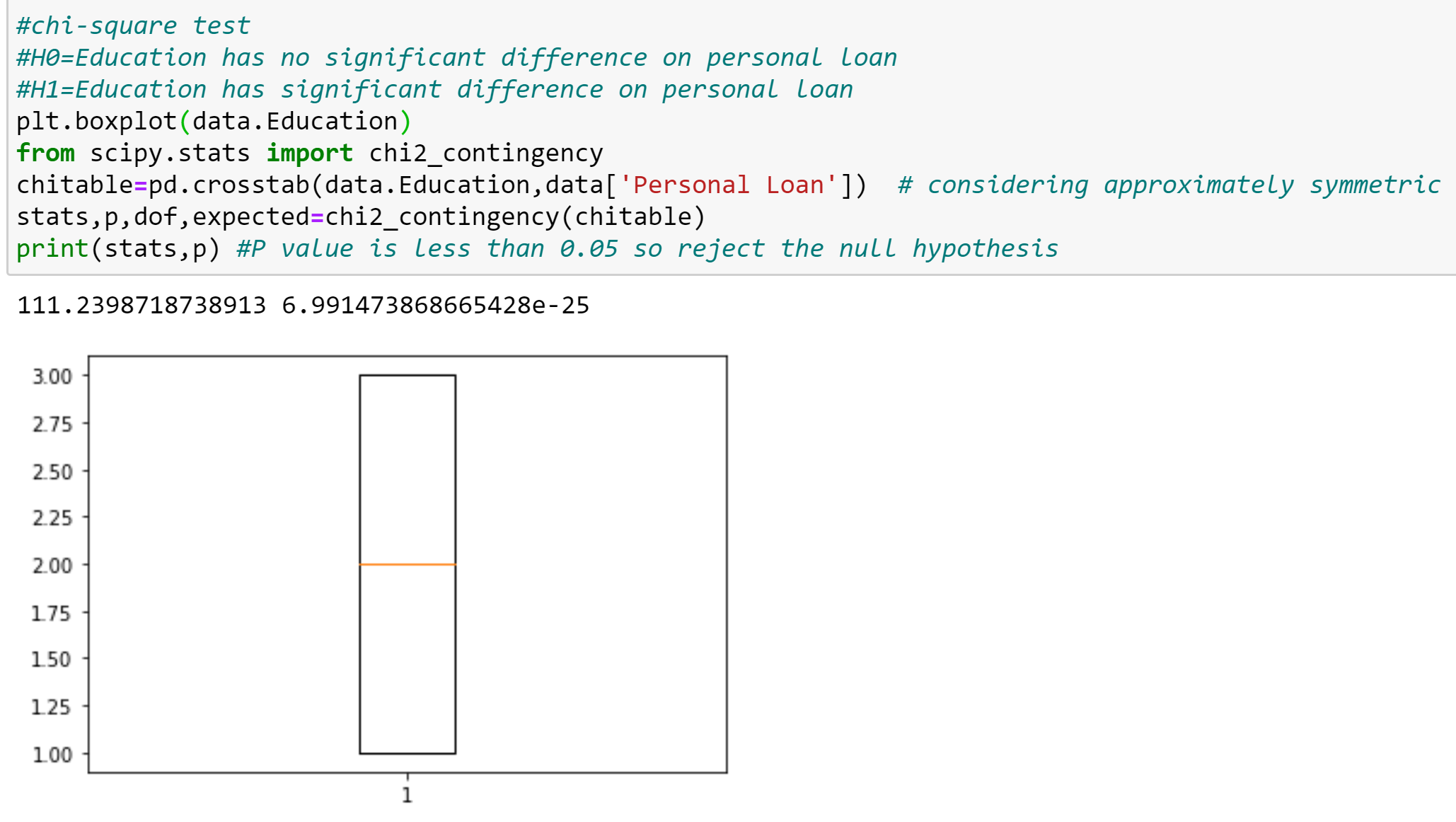
**2.** To decide the test, we need to check the distribution of education variable.

From the below plot you can see that; education is approximately has normal distribution.

Performing Chi-Square test between education and personal loan

h0= Education has no significant difference on personal loan

h1=Education has significant difference on personal loan

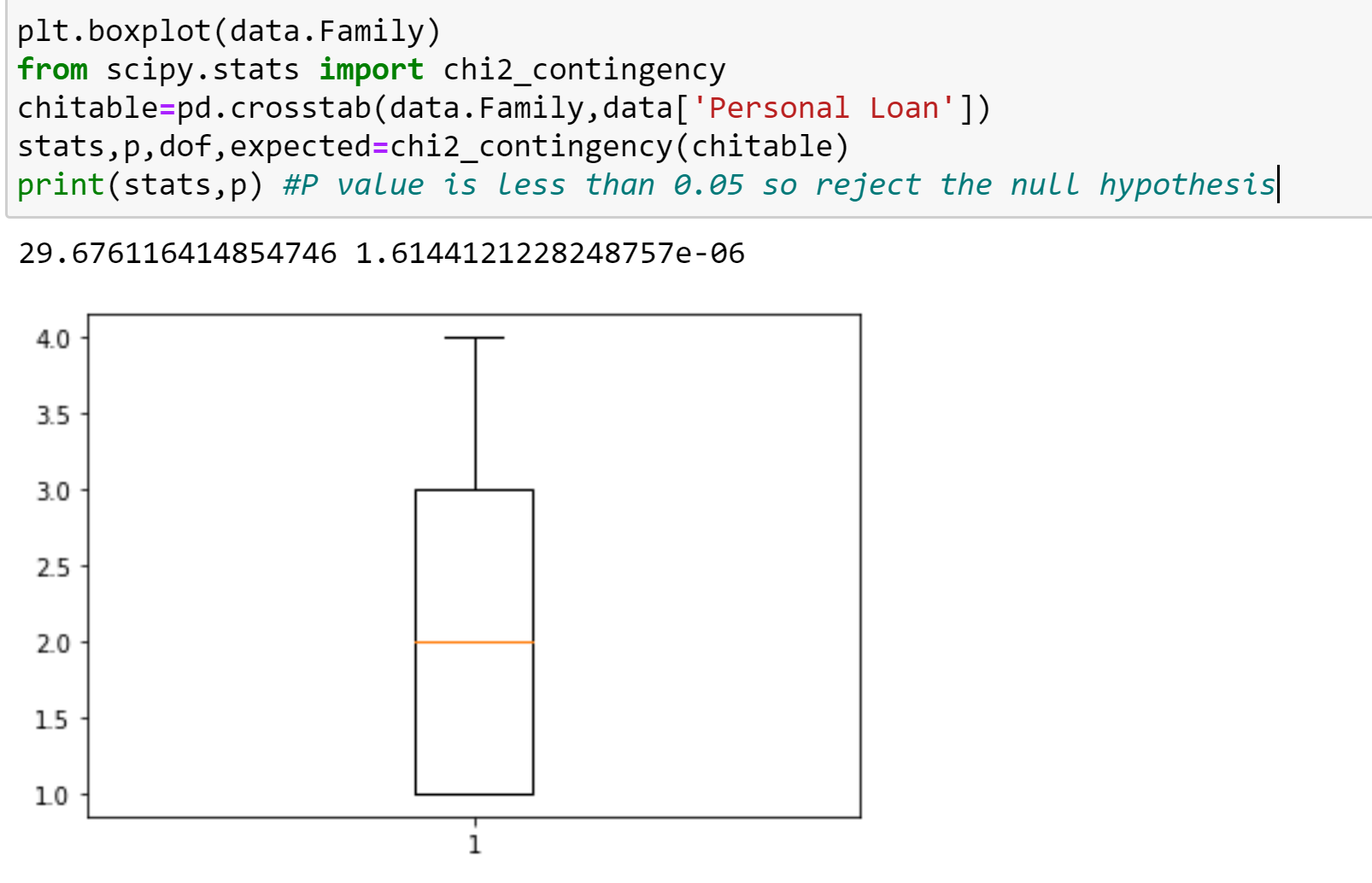


P value is less than 0.05 , so reject the null hypothesis. Education has significant effect on personal loan.

**3.**chi-square test

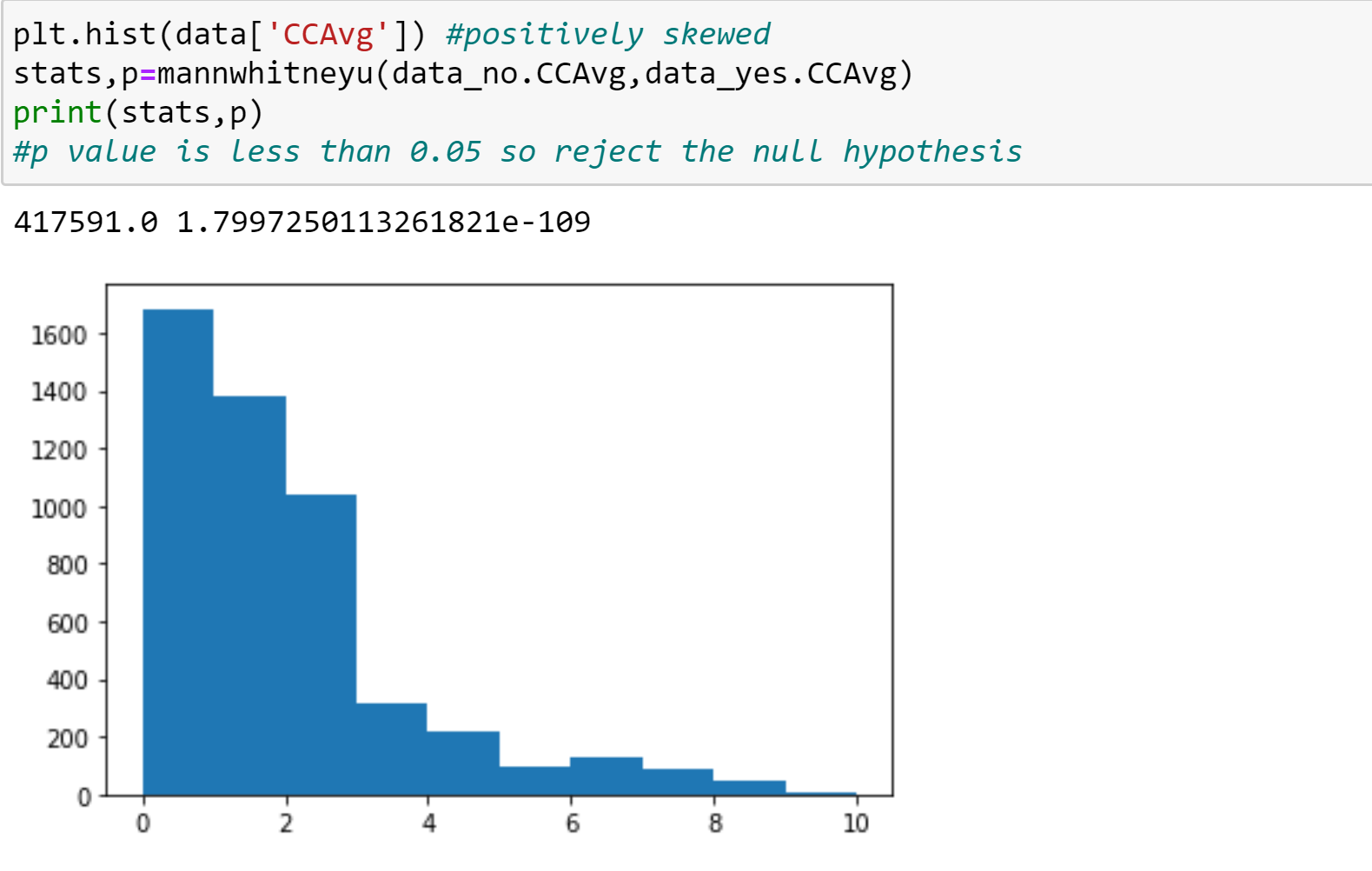
H0=Family has no significant difference on personal loan

H1=family has significant difference on personal loan

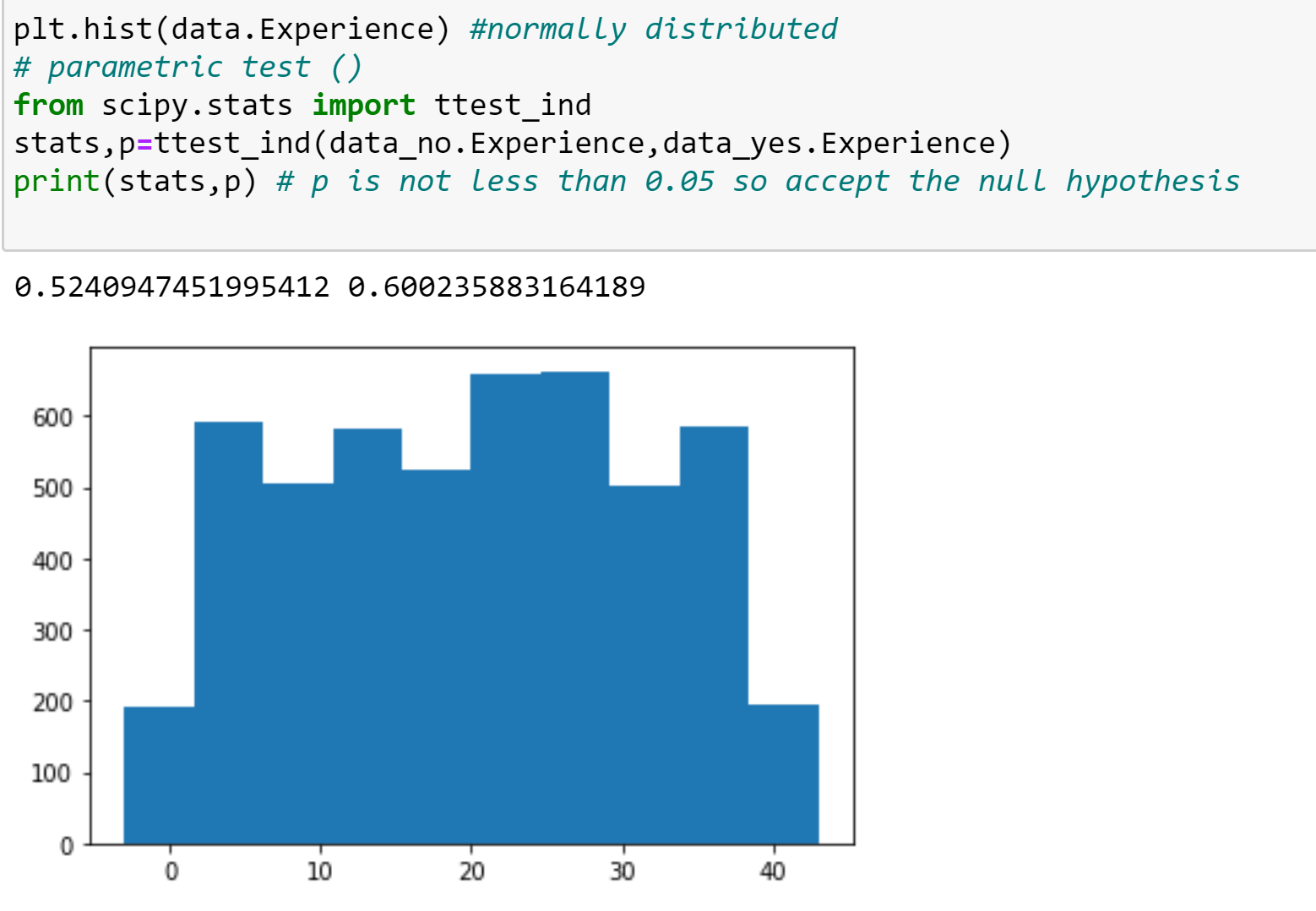


P value is less than 0.05 , so reject the null hypothesis. Family size has significant effect on personal loan.

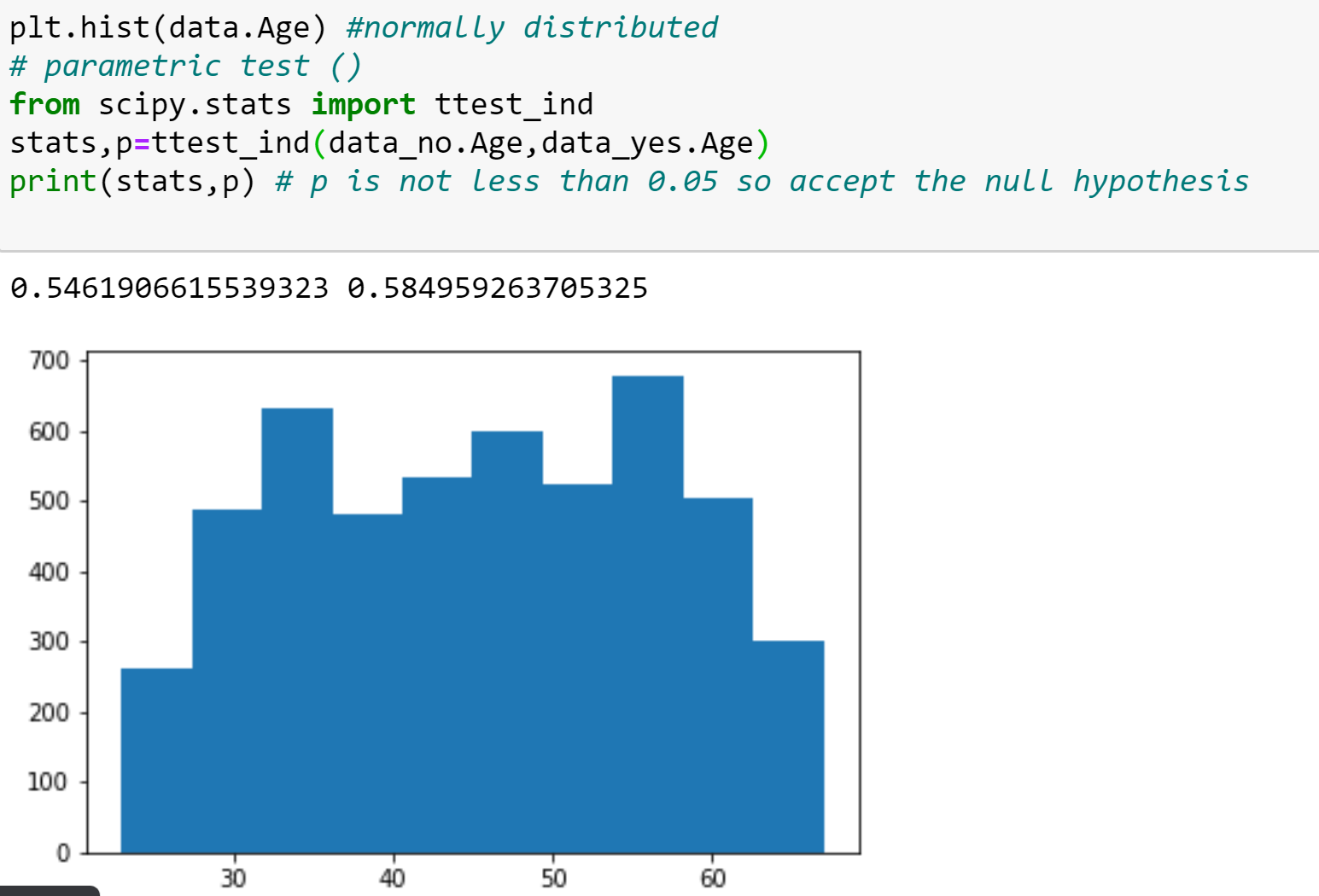
**4.** P value is less than 0.05 so reject the null hypothesis. CCAvg has significant effect on personal loan.



**5.** As data is normally distributed , 2 sample t test is applied between independent samples.P value is not less than o.o5 so accept the null hypothesis that experience has no significant effect on personal loan.



**6.** Age variable is normally distributed.But p value is not less than 0.05 so accpet the null hypothesis.Age has no significant effect on personal loan.



**Step 7: Modeling:**

As target variable is binary, we go for logistic regression.

From the below analysis: Income , family , education , online , credit card , cd account are most significant variables since p value is 0 and followed by CCAvg and securities account because p value is almost near to zero.All other variables are not significant as p value Is not less than 0.05.

Equation: p= 1/1+e-(B0+B1X1+B2X2…..)

By substituting the coefficient and intercept values with the sample data or test data we can obtain the value of p.

If p is greater than 0.5, then personal loan is approved by the bank else it is rejected.

