ALLLIFE BANK PROJECT 4

Sarah A. Thomas

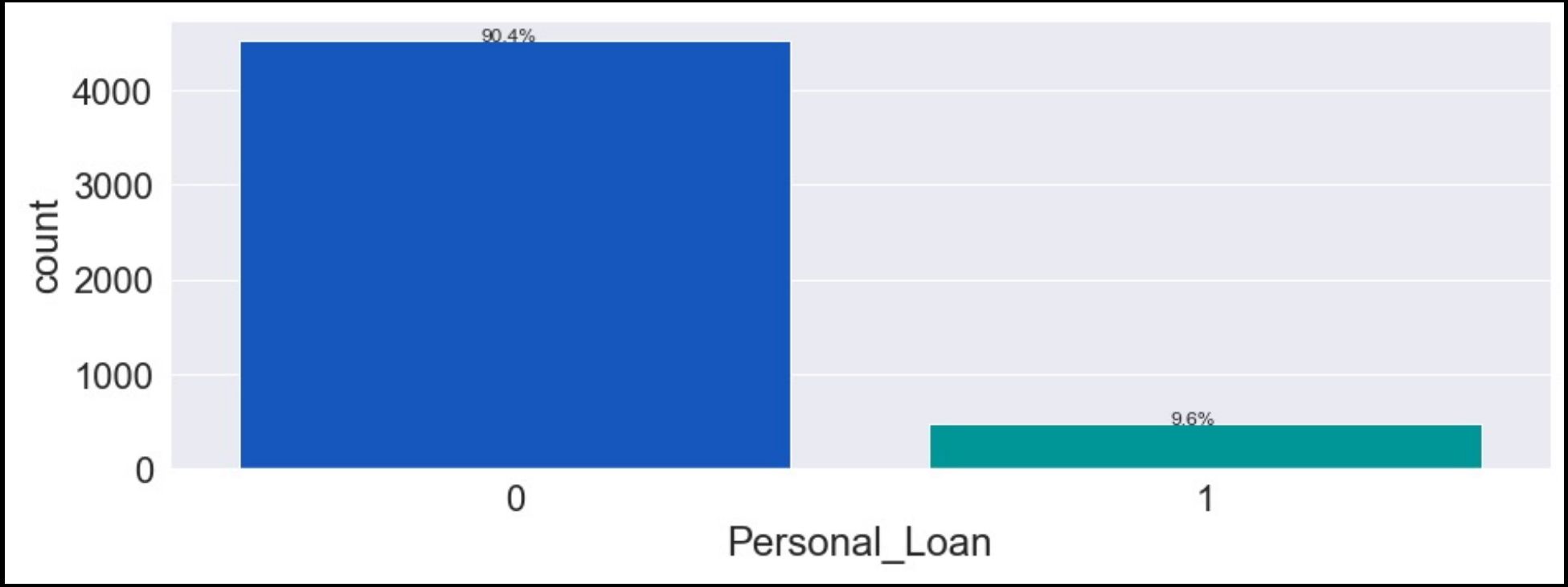
BUSINESS PROBLEM OVERVIEW AND SOLUTION APPROACH

- Currently the growing customer base is primarily made up of liability customers (depositors).
- AllLife Bank wishes to grow the number of customers who borrow (asset customers) in order to earn money through interest on loans.
- Last year, a campaign was run that converted 9% of those offered loans to asset customers. AllLife Bank wishes to build on this success.
- The current task for the Data Science department is to design a model that will predict potential customers who are likely to accept a loan if offered. Statistically, the approach will be to maximize the Precision metric. We will want to minimize the number of people we shut out of accepting a loan. In other words, we want to minimize predicting they'll say "no" when they'll actually say "yes."

DATA OVERVIEW

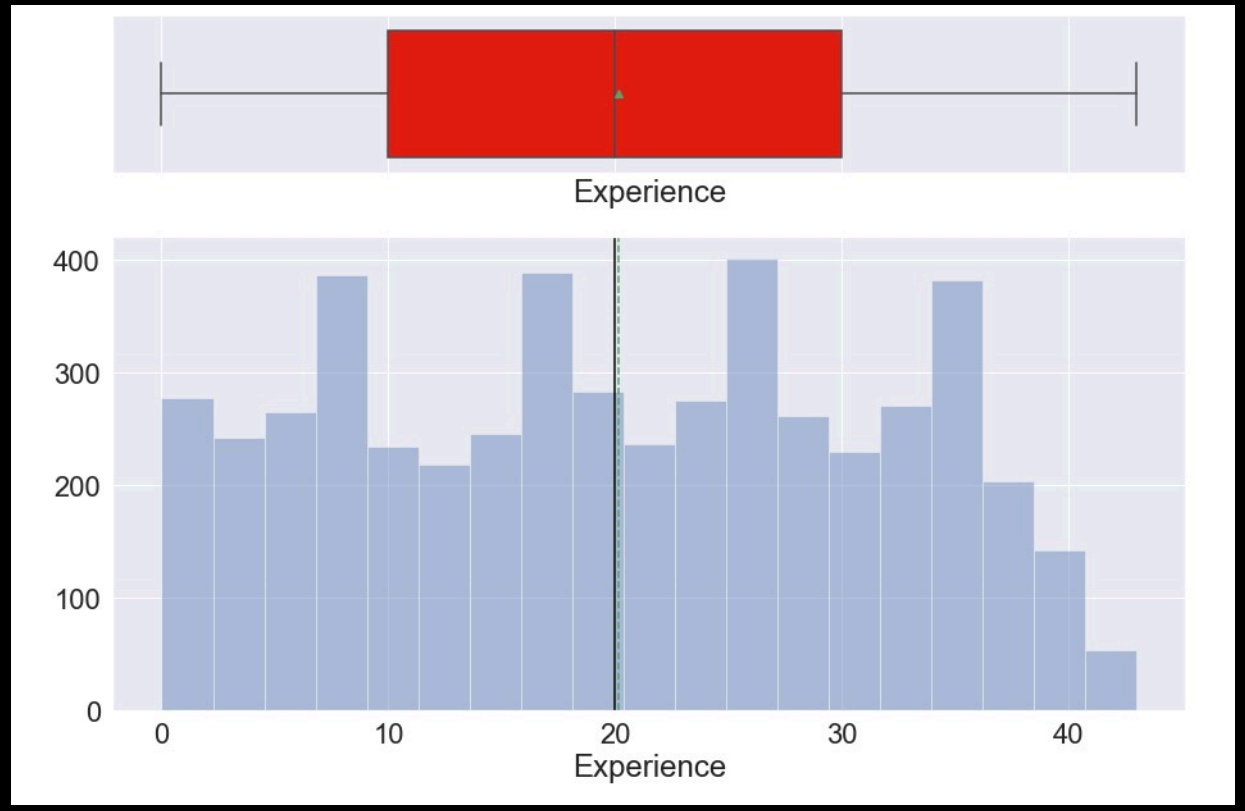
Variable	Description
ID	Customer ID
Age	Customer's age in years
Experience	Number of years of professional experience
Income	Annual income of customer (in thousands of dollars)
ZIP Code	Home Address ZIP code
Family	Customer's family size
CCAvg	Average monthly spending on credit cards (in thousands of dollars)
Education	Education Level: 1-Undegraduate, 2-Graduate, 3-Advanced/Professional
Mortgage	Value of house mortgage (if any)
Personal_Loan	(Target Variable) Did customer accept loan offered in last campaign? (1-Yes, 0-No)
Securities_Account	Does the customer have a securities account with AllLife? (1-Yes, 0-No)
CD_Account	Does the customer have a CD account with AllLife (1-Yes, 0-No)
Online	Does the customer use online banking? (1-Yes, 0-No)
CreditCard	Does the customer use a credit card issued by another bank (excluding AllLife)? (1-Yes, 0-No)

EDA – PERSONAL LOAN



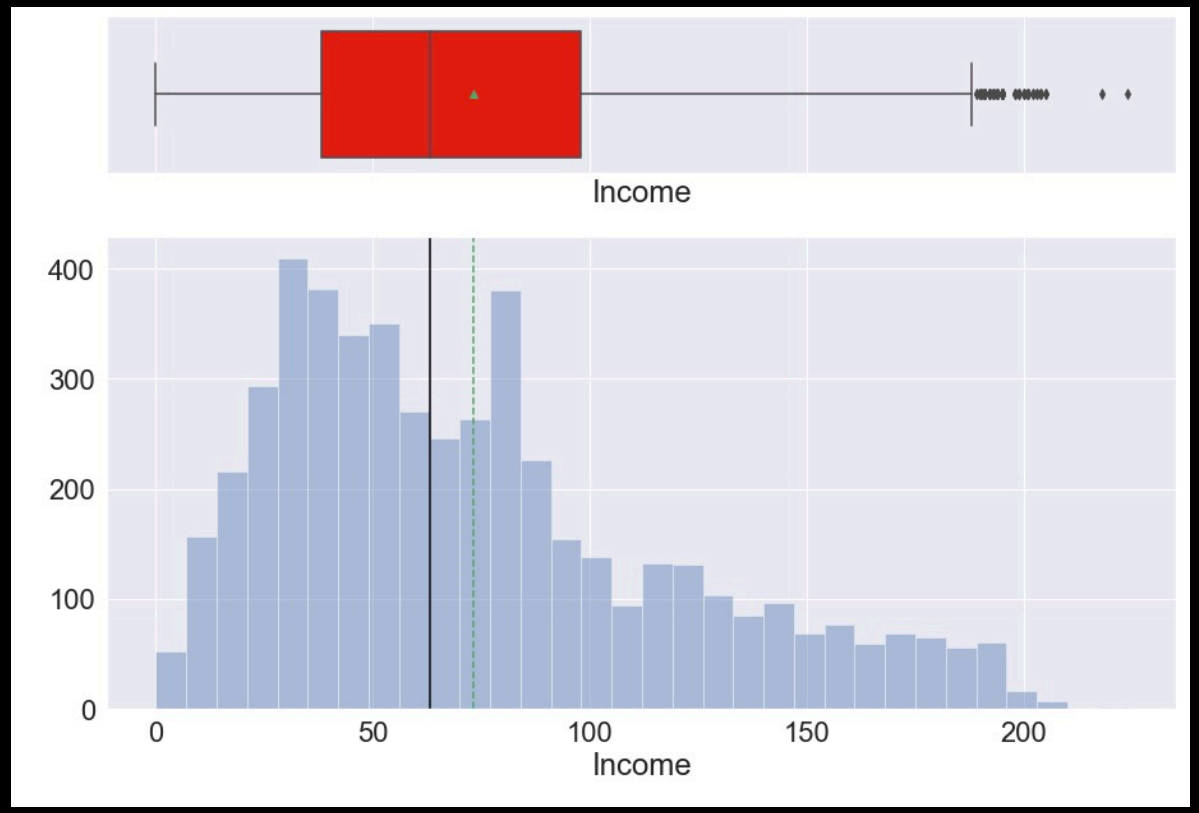
90.4% did not accept the personal loan offered in the last campaign.

EDA - EXPERIENCE



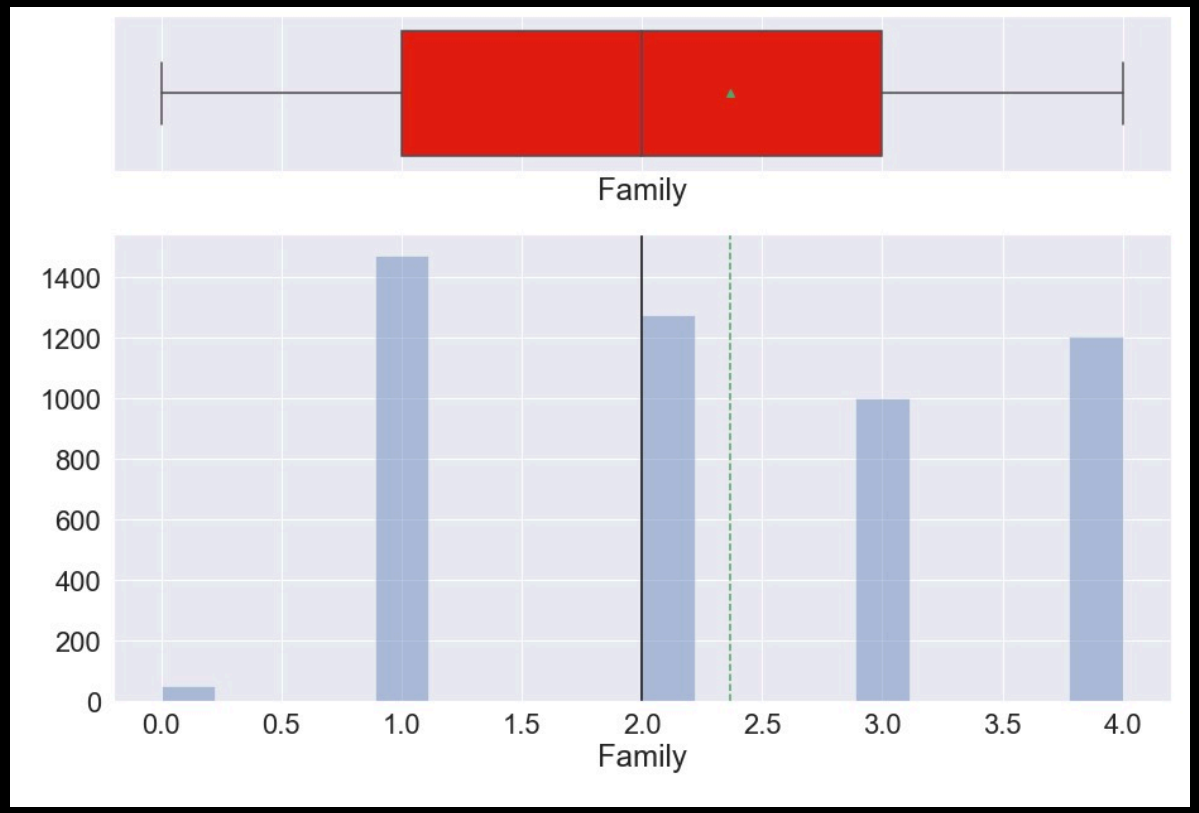
Experience has a symmetrical distribution. Negative values were changed to 0.

EDA - INCOME



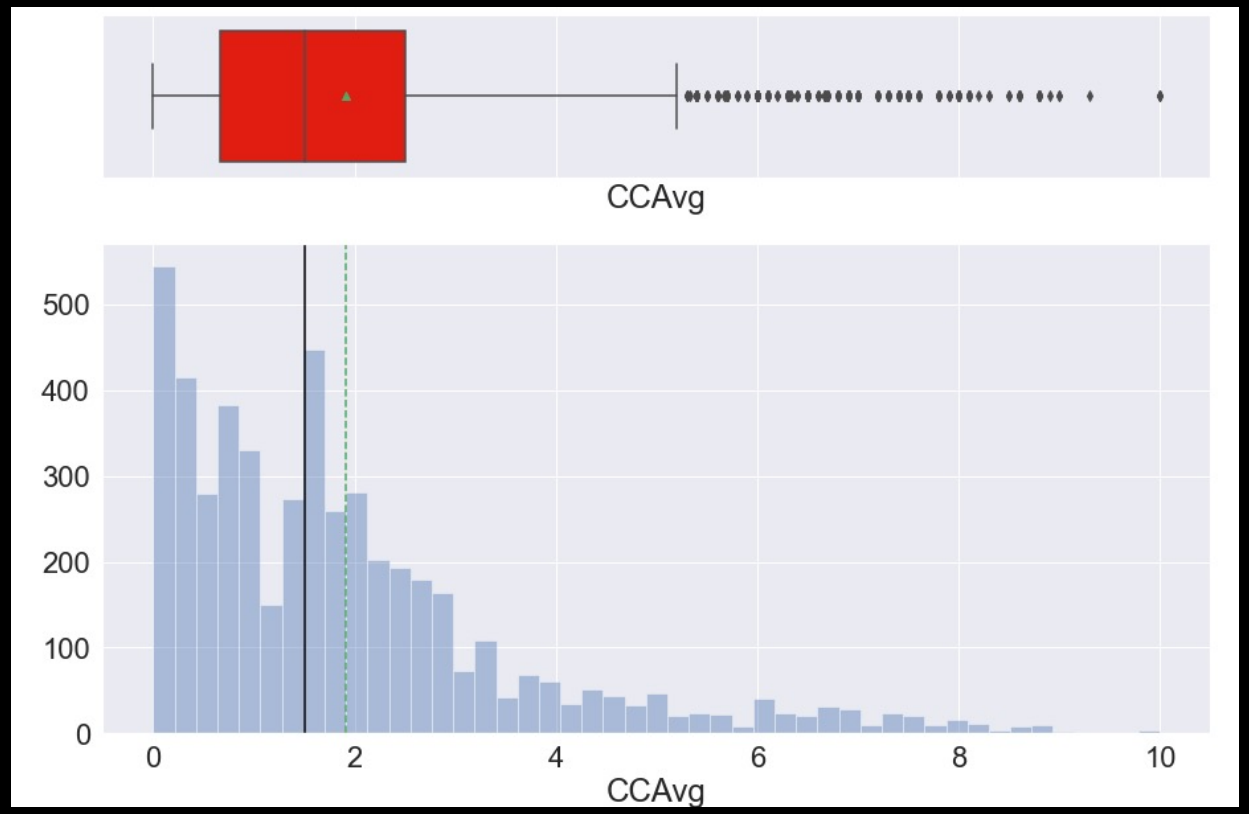
The mean (73.77) is greater than the median (64.0). Income is right-skewed. Outliers were treated.

EDA - FAMILY



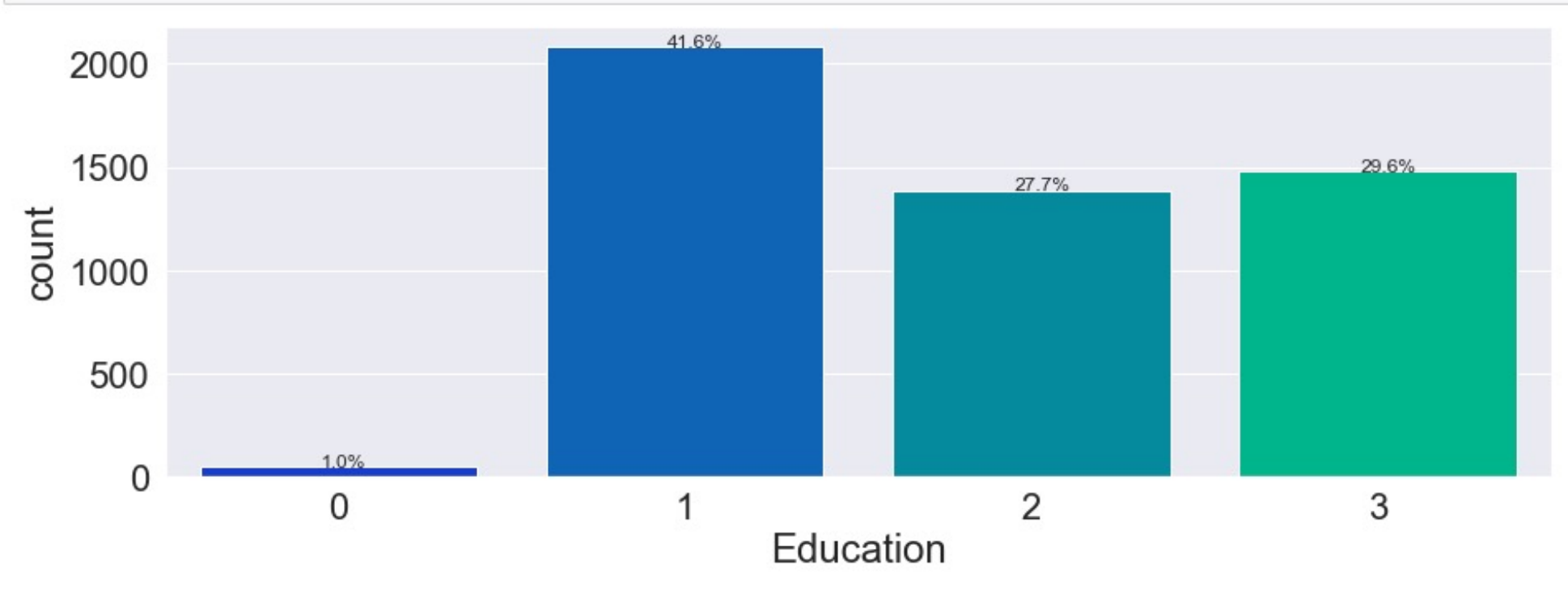
The mean (2.4) is close to the median (2.0) indicating symmetrical distribution. Values of “0” for family were changes to 1.

EDA - CCAVG



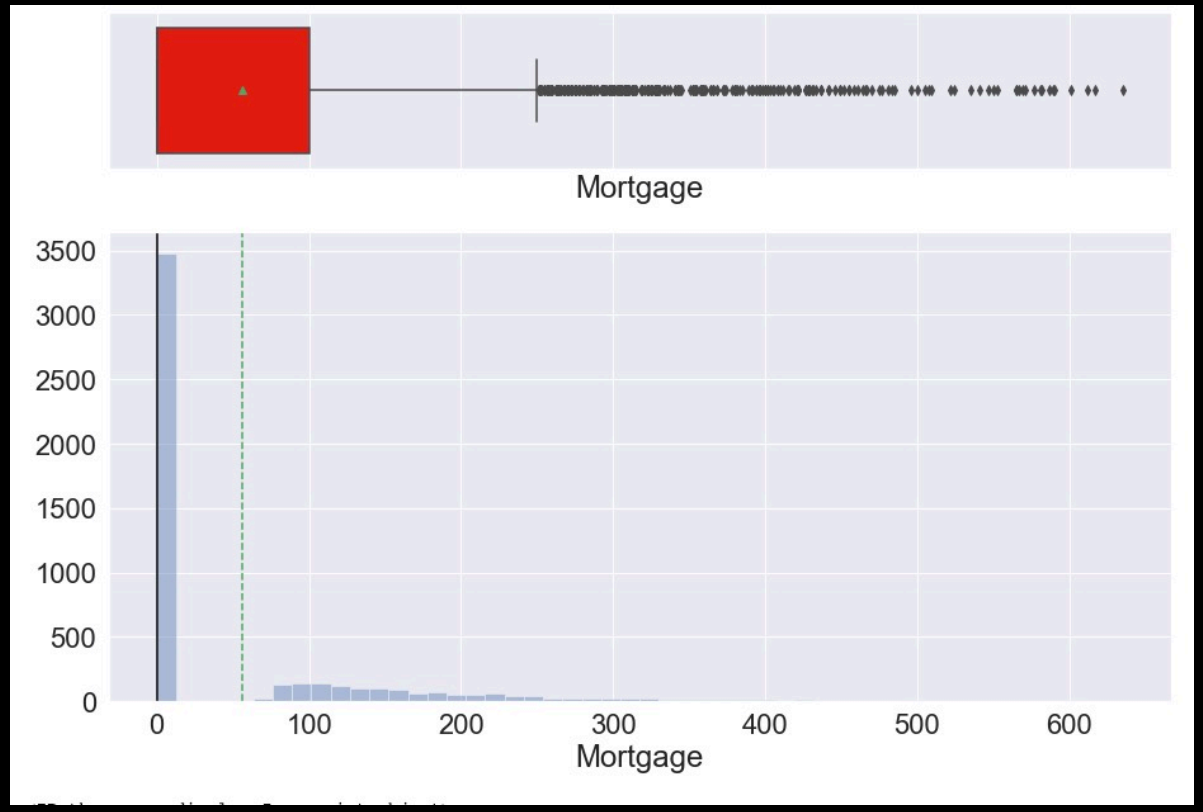
The mean (1.94) is greater than the median (1.5). CCAvg is right-skewed. Outliers were treated.

EDA - EDUCATION



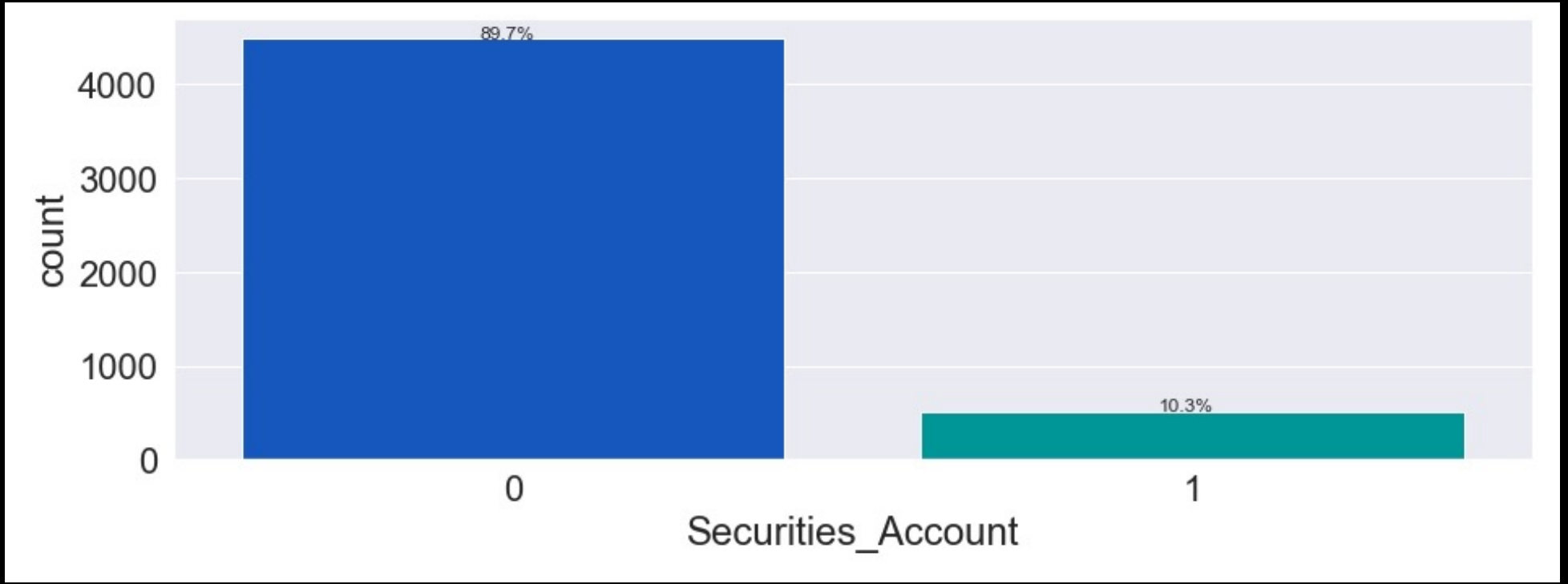
Most customers have at least an undergraduate degree (41.6%). 1.0% do not have a degree.

EDA - MORTGAGE



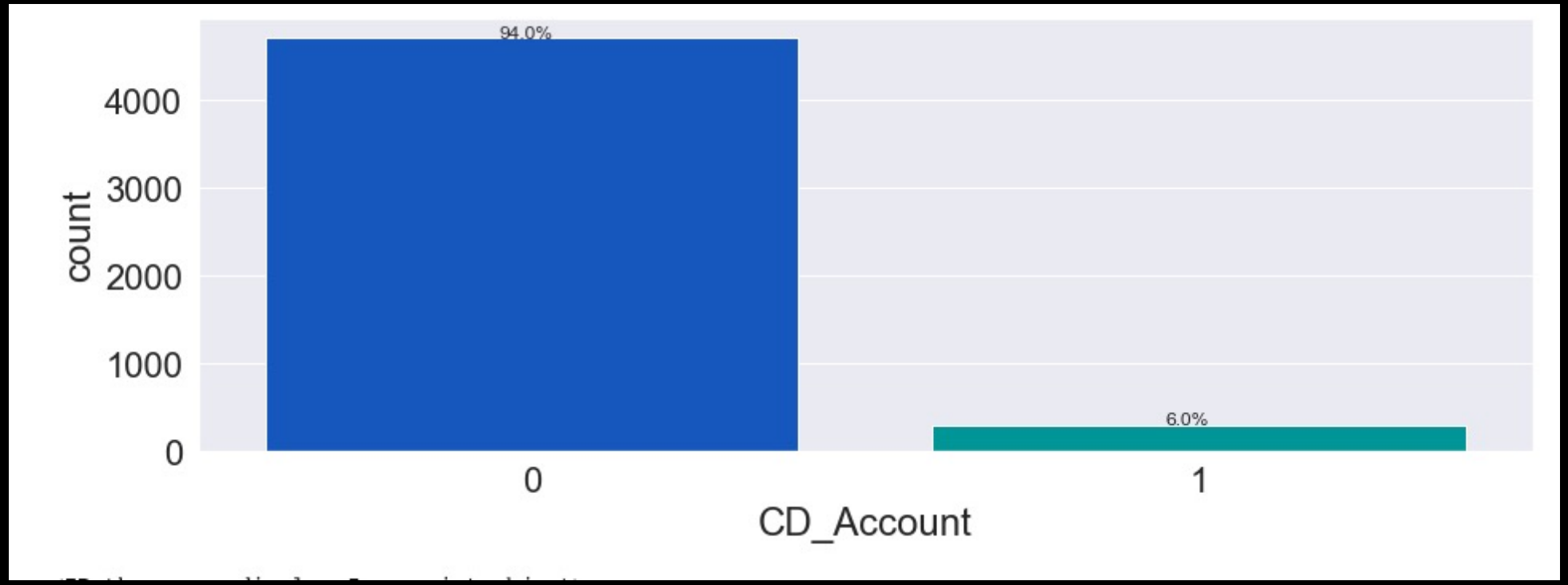
Mortgage is right-skewed. Outliers were treated.

EDA – SECURITIES_ACCOUNT



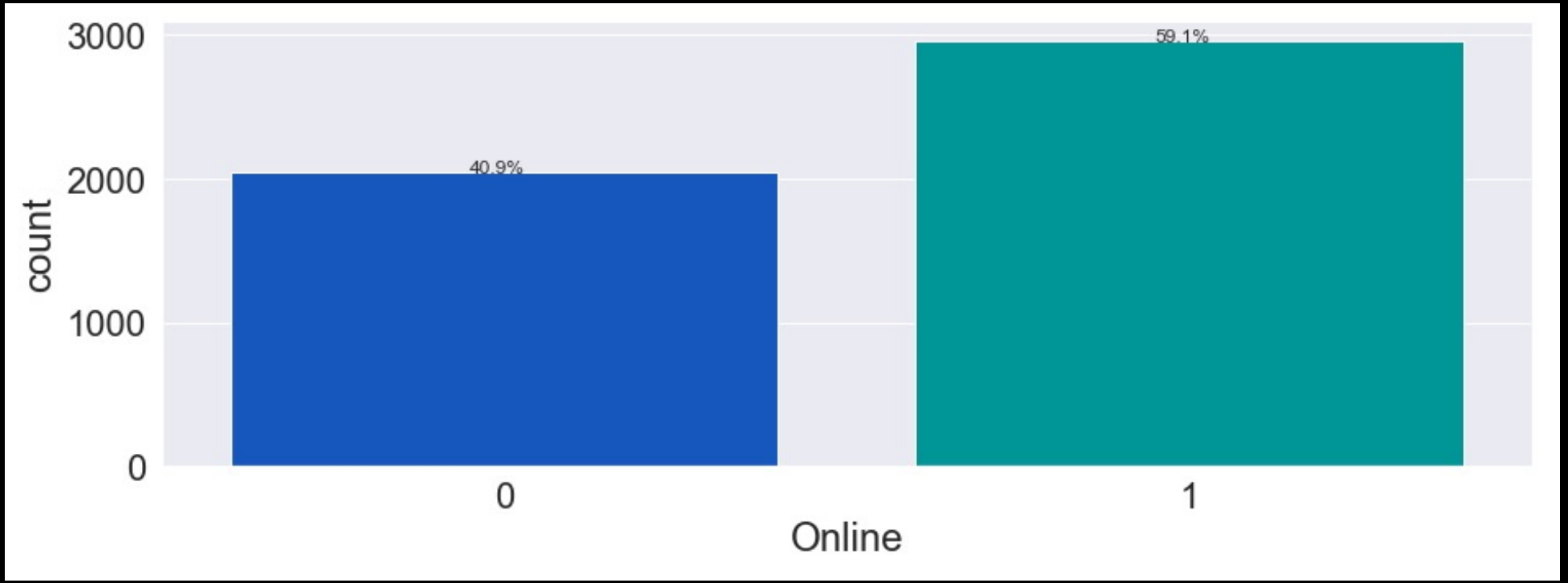
Most customers do not have a securities account (89.7%).

EDA – CD_ACCOUNT



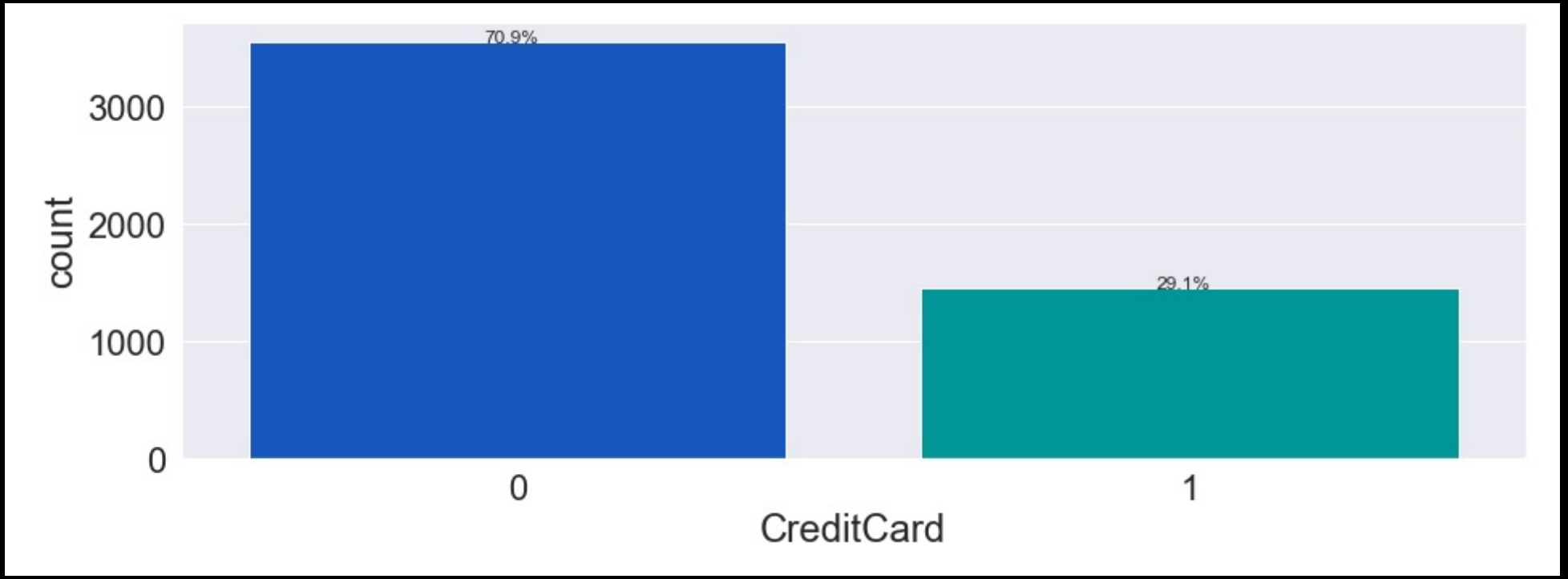
Most customers do not have a CD account (94%).

EDA - ONLINE



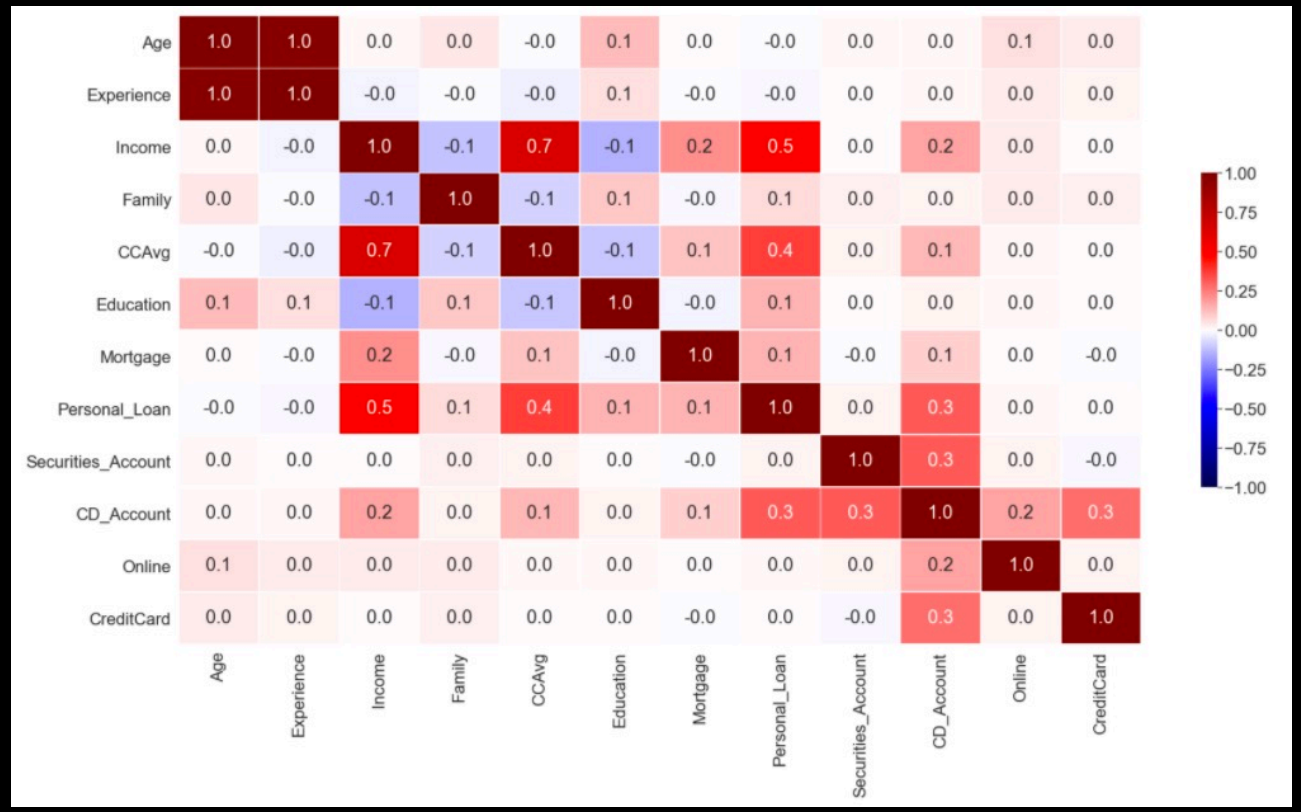
Most customers use online banking (59.1%).

EDA - CREDITCARD



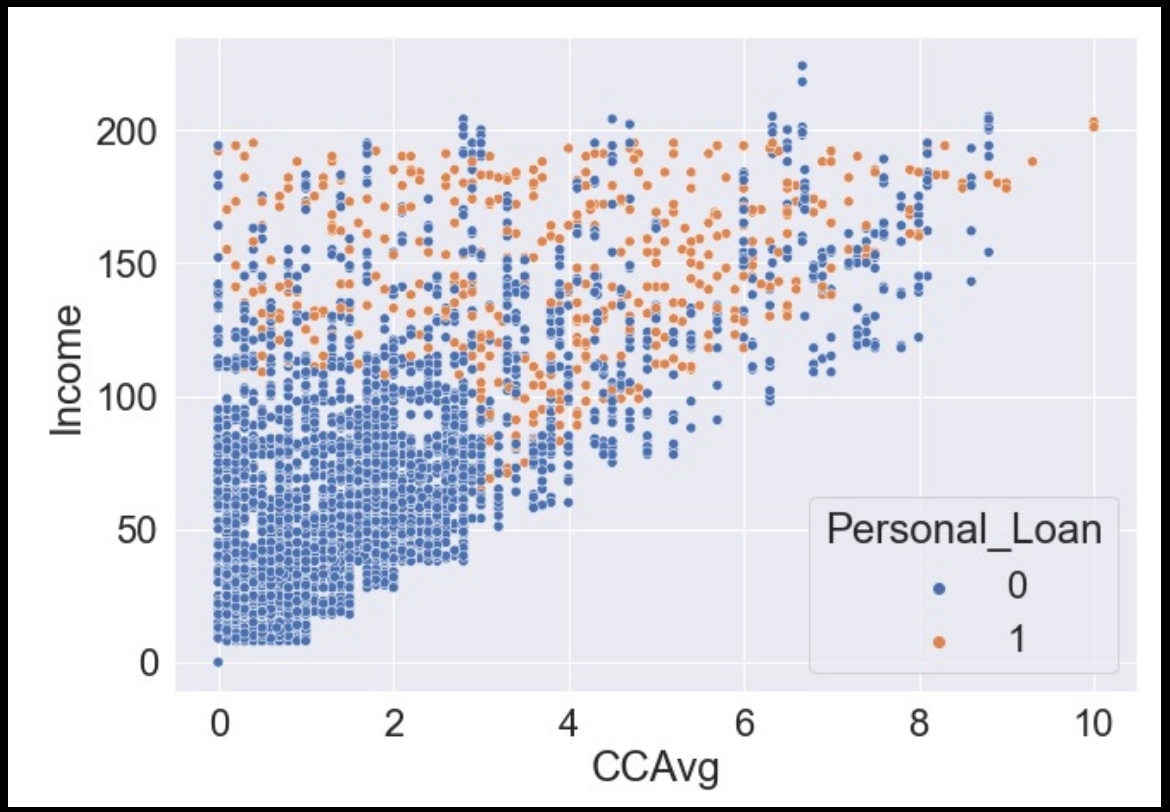
Most customers do not use a credit card from another bank (70.9%).

CORRELATION



- Age and Experience are perfectly correlated.
- High correlation exists between Income and CCAvg.
- Moderate correlation exists between Personal_Loan and Income, Personal_Loan and CCAvg.
- Slight correlation exists between Income and CD_Account, Income and Mortgage, Personal_Loan and CD_Account, Securities_Account and CD_Account, CD_Account and CreditCard, CD_Account and Online.

MULTIVARIATE ANALYSIS

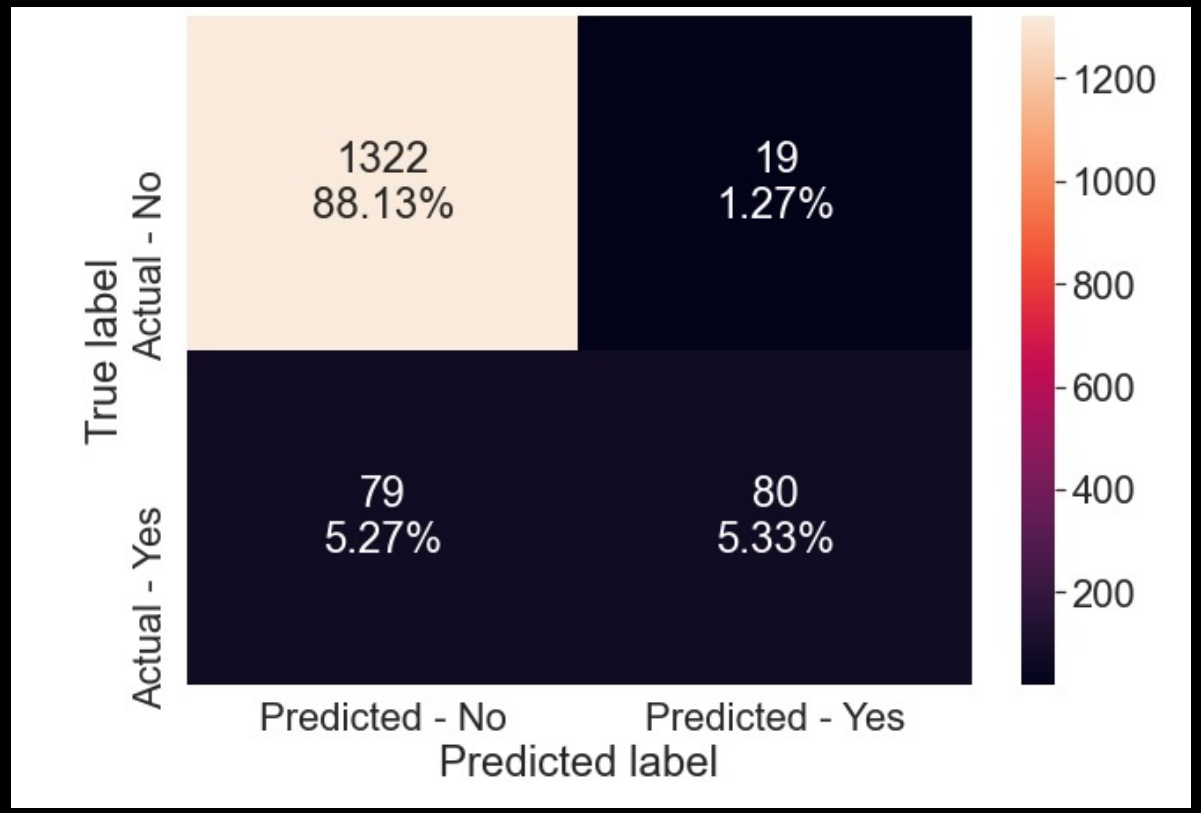


Higher income and higher monthly spending on credit card seems to correspond with acceptance of personal loan.

DROPPED VARIABLES

- Age and Experience are perfectly correlated. Since they are telling the same story, age was dropped.
- Zip codes can bias the dataset in a discriminatory way (e.g., serve as proxy for race) and so was dropped.
- ID was irrelevant and thus dropped from the data.

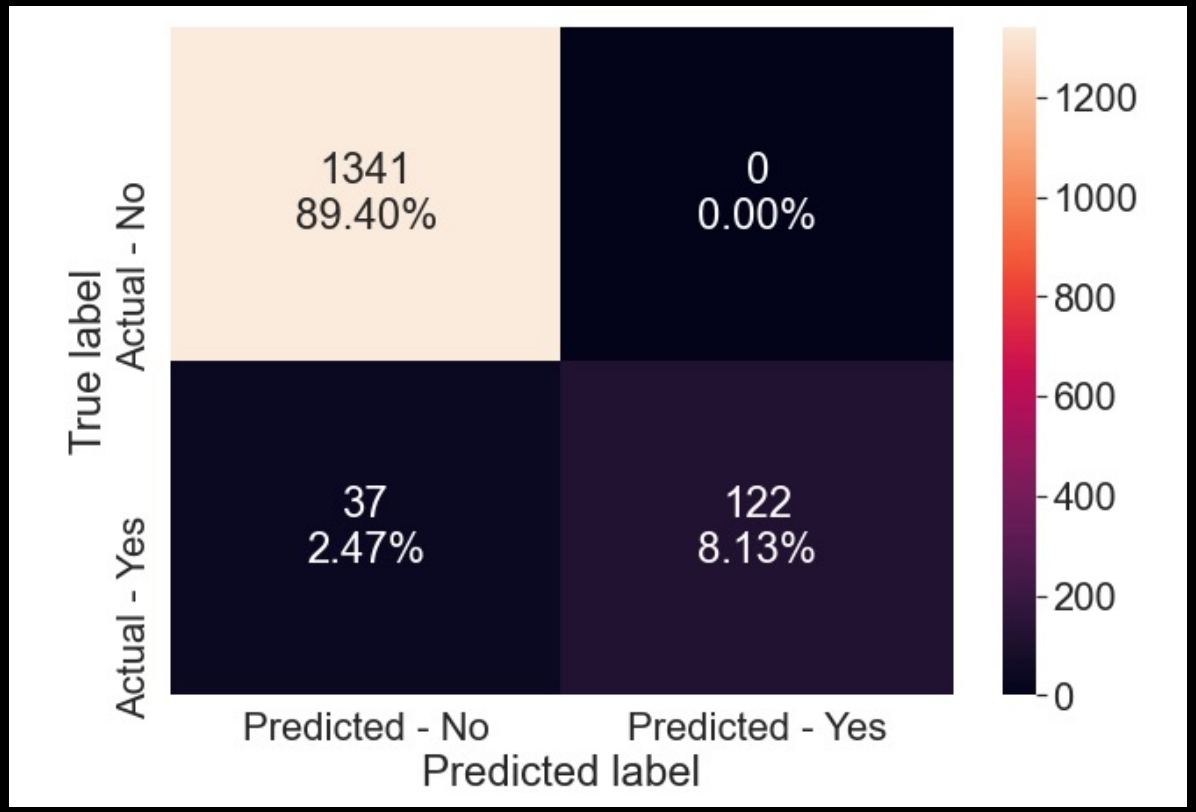
LOGISTIC REGRESSION – SKLEARN



- Accuracy on training set: 0.93
- Accuracy on test set: 0.93
- Recall on training set: 0.47
- Recall on test set: 0.50
- * Precision on training set: 0.78
- * Precision on test set: 0.81
- F1 on training set: 0.58
- F1 on test set: 0.62

A model was created using statsmodels but precision results were not as satisfying.

DECISION TREE



- Accuracy on training set: 0.98
- Accuracy on test set: 0.98
- Recall on training set: 0.83
- Recall on test set: 0.77
- * Precision on training set: 1.00
- * Precision on test set: 1.00
- F1 on training set: 0.91
- F1 on test set: 0.87

Before pruning, tree was more accurate but less precise. It was also much more complex. This has a depth of 3.0.

CONCLUSION

- Income, Education_1, Family, and CCAvg (in that order) are the most important variables in determining whether a customer will accept a loan if offered.
- The average customer is 45 years old, has an undergraduate education, is in a two-member family, and spends $1500 per month on their credit card(s).
- Higher income and higher monthly spending on credit card seems to correspond with acceptance of personal loan.

Comparison of all models created

	Model	Train_Precision	Test_Precision
0	LR with SKLearn	0.78	0.81
1	LR with statsmodels	0.77	0.80
2	Decision tree with pre-pruning	1.00	0.91
3	Decision tree with post-pruning	1.00	1.00