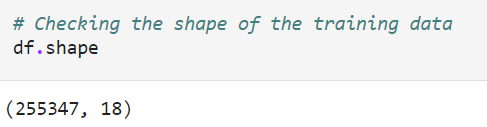
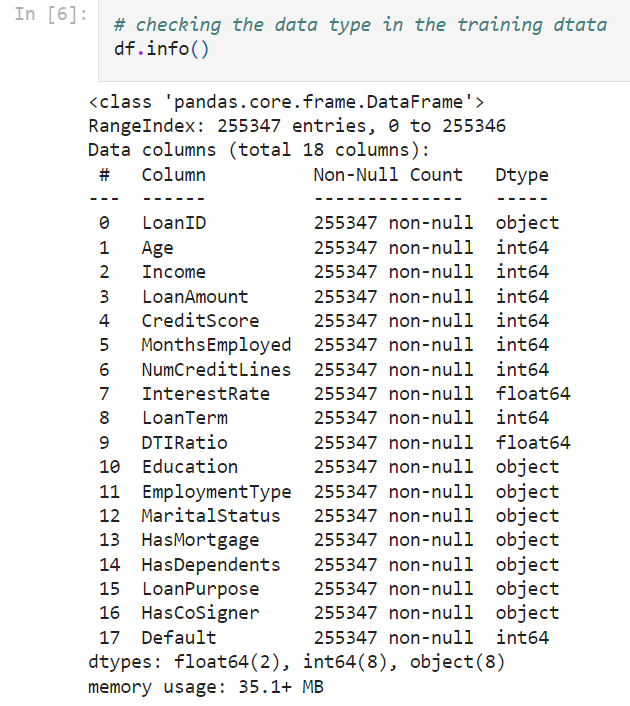
***LOAN\_DEFAULT PREDICTION***

***Data Structure***

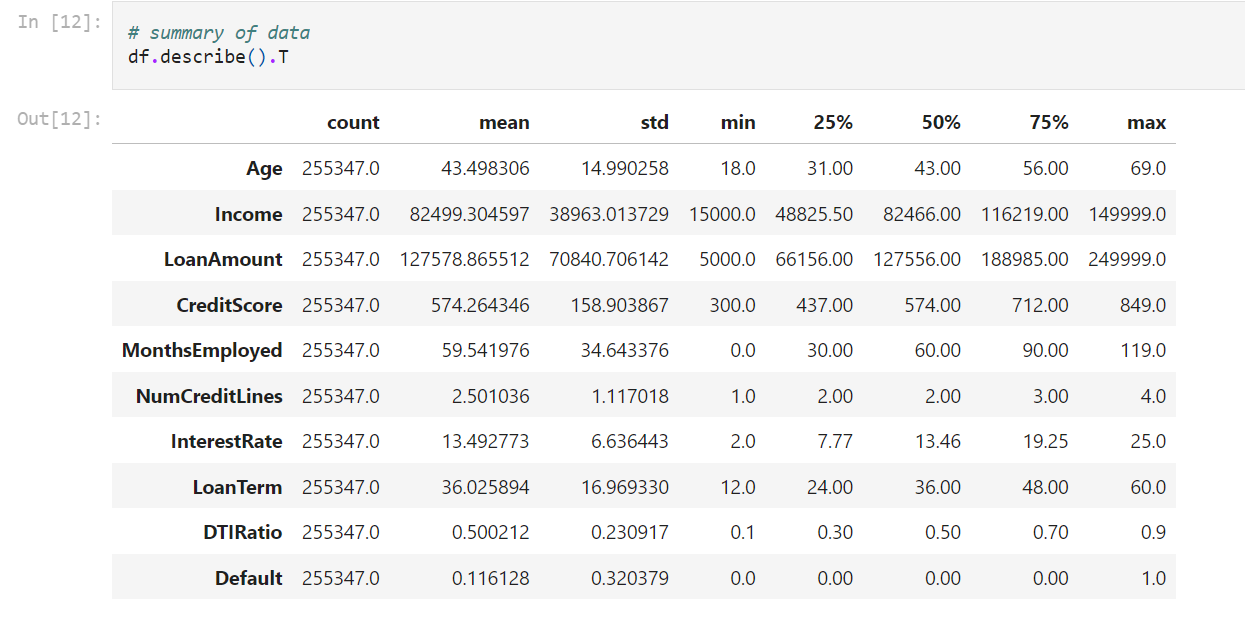
Lets first learn about the dataset that we are working upon.



As we can see from the above picture, our dataset consists of 18 columns and 255347 rows in total.



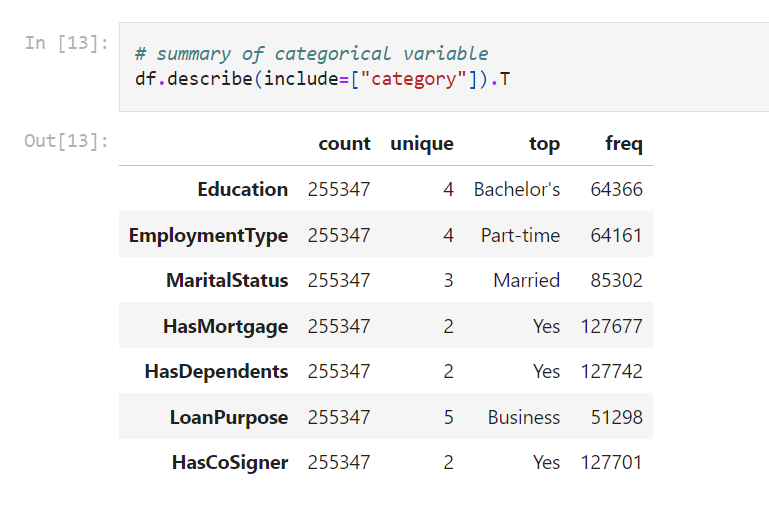
Out of the total 18 variables, some are numerical and some are categorical variables which has a object datatype(we convert it into category datatype in future). We can also see that there aren’t any missing values in our dataset and there aren’t any null values too.



Here we are looking at the statistical summary of the numerical variables.

* Age- The mean age of the dataset is around 43 years to 44 years with a standard deviation of 14.99. The minimum age of a customer who has applied for loan is 18 years and the maximum being 69 years.
* Income- On an average the customers have a mean income of 82499 units with the minimum income being 15000 units and the maximum being 149999 units.
* Loan Amount- The average loan amount that customers have demanded is 127578 units which is a bit higher than the average income. The lowest and the highest loan amount that customers have demanded are 5000 and 249999 units respectively.
* Credit Score- On an average the customers have a credit score of 574, with the lowest being 300(maybe of students) and the highest being 849.
* Months Employed- On an average customers have a work experience of almost 5 years but this data has a low of 0 months making it clear that students have also applied for loan in past, maybe an education or a personal loan.
* Number of credit lines- On an average customers have 2 credit lines, the minimum being 1 and the maximum being 4.
* Interest Rate- On an average across various loans the interest rate is 13.5%. The lowest interest rate is 2% which might be given on an education loan and the highest interest rate being 25%.
* Loan Term- On an average customers take a loan for a period of 3 years, the lowest being a year and the highest being 5 years.
* DTI Ratio- The debt to income ratio stands to be 0.5 on average with the lowest being 0.1 and the highest being 0.9.

From this table we can see the data is balanced with minimum to no outliers since mean is almost equal to median for most of the cases which shows a normally distributed data.

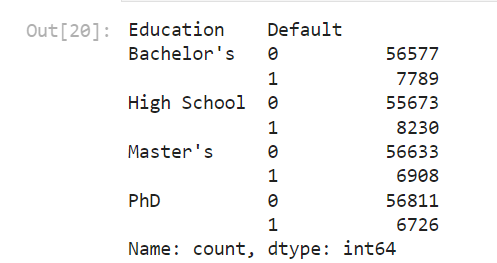


This table shows the summary of the categorical variables.

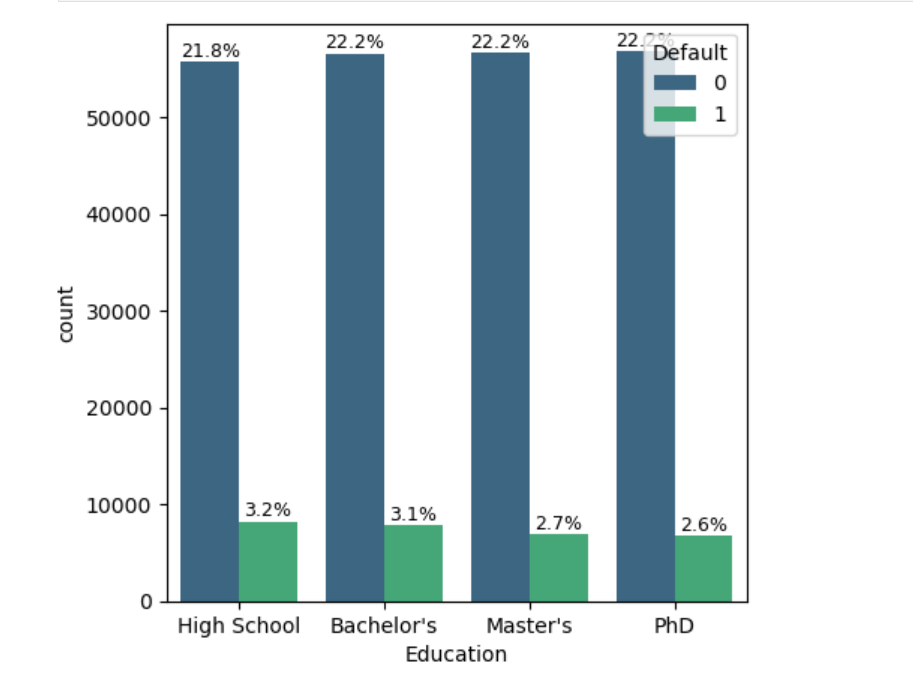
* Most of the customers have educational qualification till bachelors, having a part time job, married, has a mortgage, has dependants and needs the loan to start a business.
* Most of them also have a cosigner maybe since most of them are married or has dependants.

***EDA***

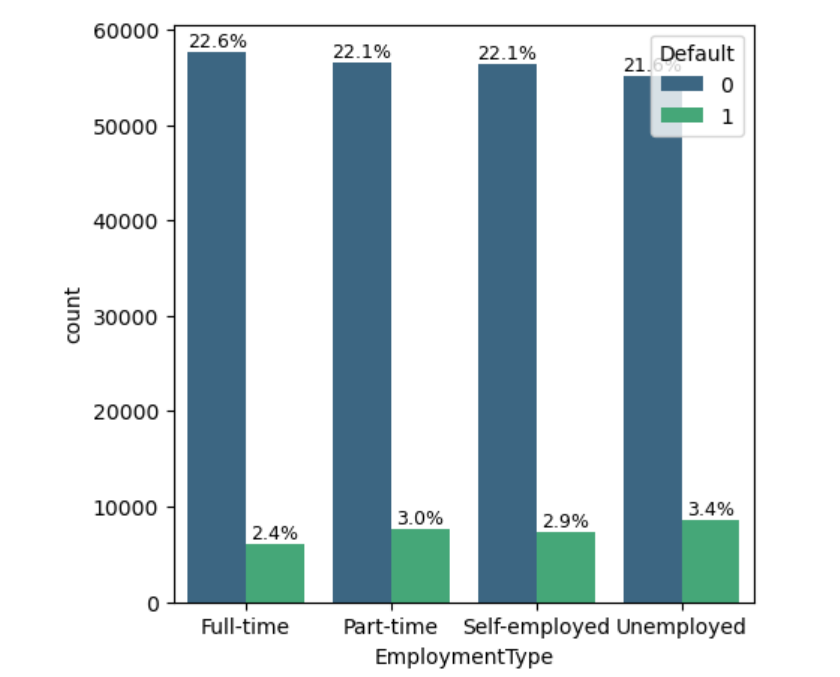
This section shows some relationships between the variables and whether a customer will default or not



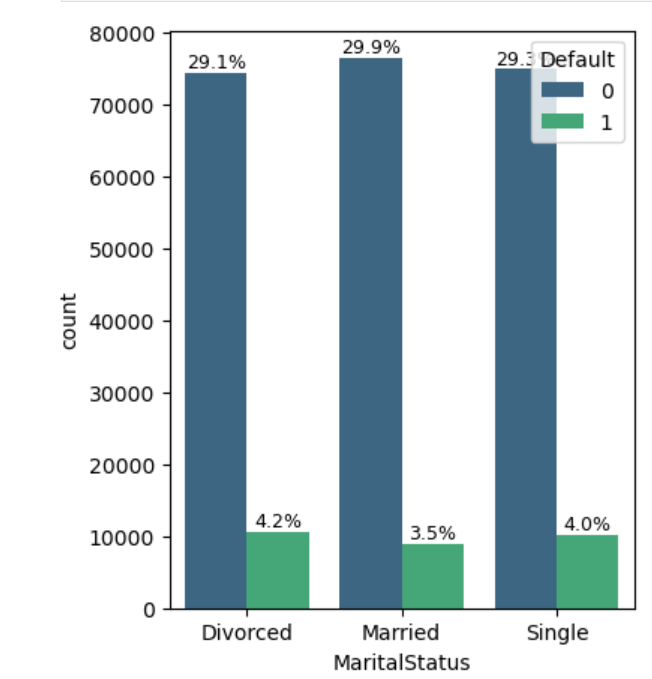
This table shows that people who have educational qualification till highschool have defaulted the most followed by bachelors, masters and PhD. This shows that a person who has a higher educational qualification has lesser chances of defaulting the loan.



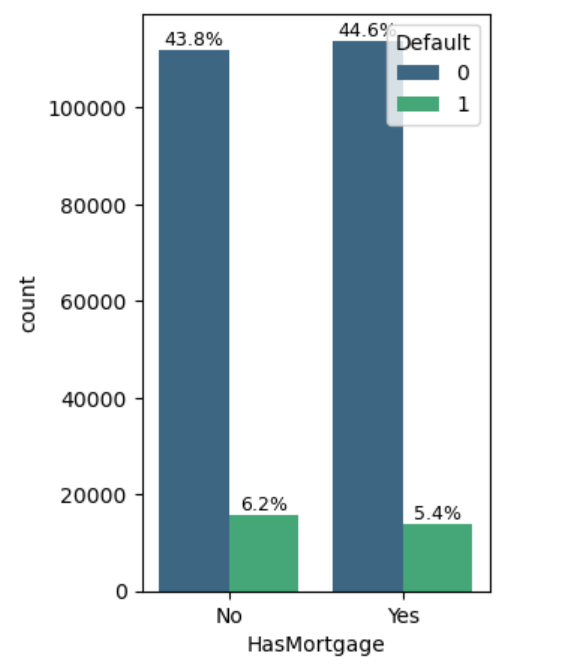
The same can be seen from the graph above.



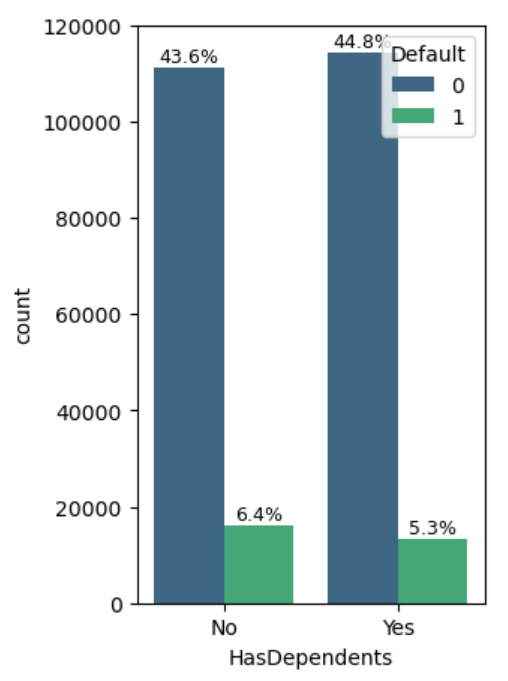
From the above graph we can see that a person who is unemployed has the most chances of defaulting a loan followed by being part time employed, self employed and full time employed. This might give us an idea that people who have educational qualification till high school and unemployed have most chances of defaulting a loan.



This graph shows the relationship between marital status and defaulting loan. People who are divorced have the most chances of defaulting followed by singles and married people.

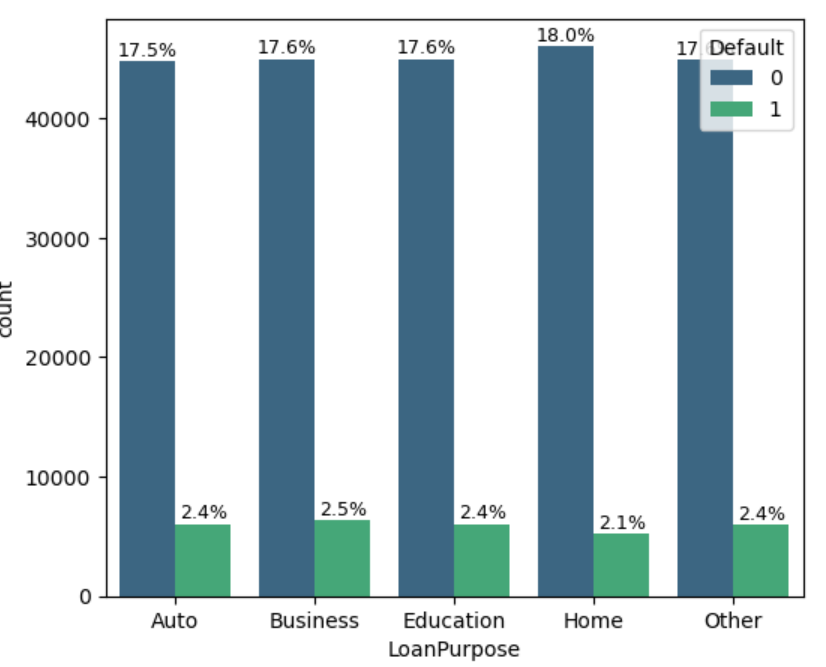


The above chart shows that people who doesnot have mortgage are more likely to default as compared to people who have mortgage.

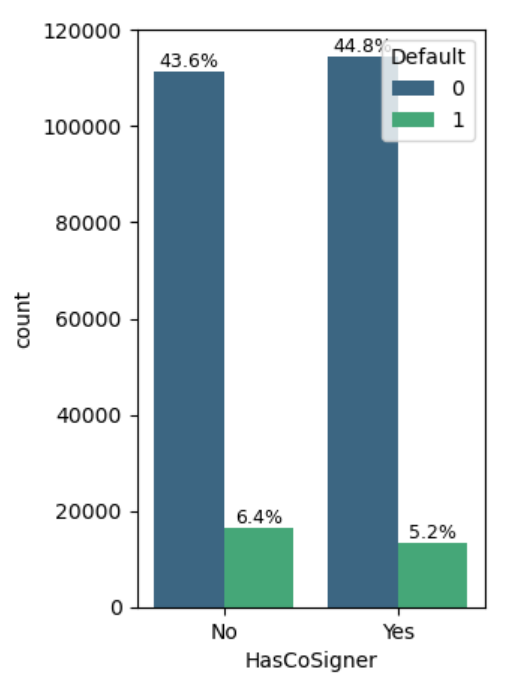


This graph shows that people who doesnot have any dependants are more likely to default a loan as compared to people who has dependants.

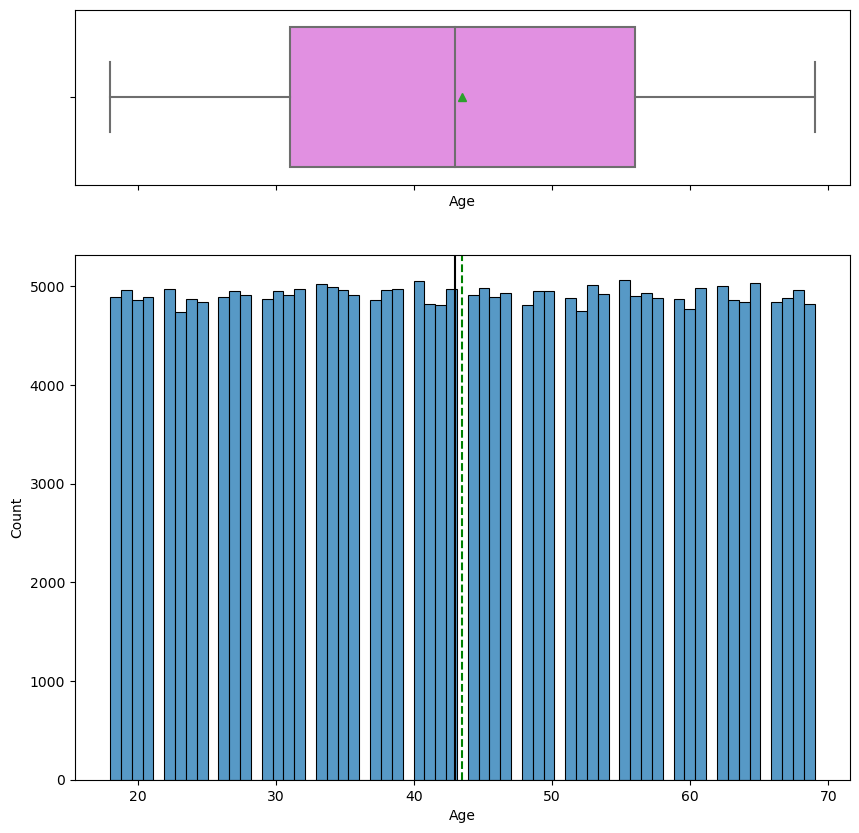
From the above graphs we can see that people who have educational qualification till high school, are unemployed, single,has no mortgage and has no dependants are likely to default the loan. These characteristics can be seen in a student who wants to take a loan to further studies after high school. So educational loan might have the most default rates. This theory can be confirmed by our next graph.



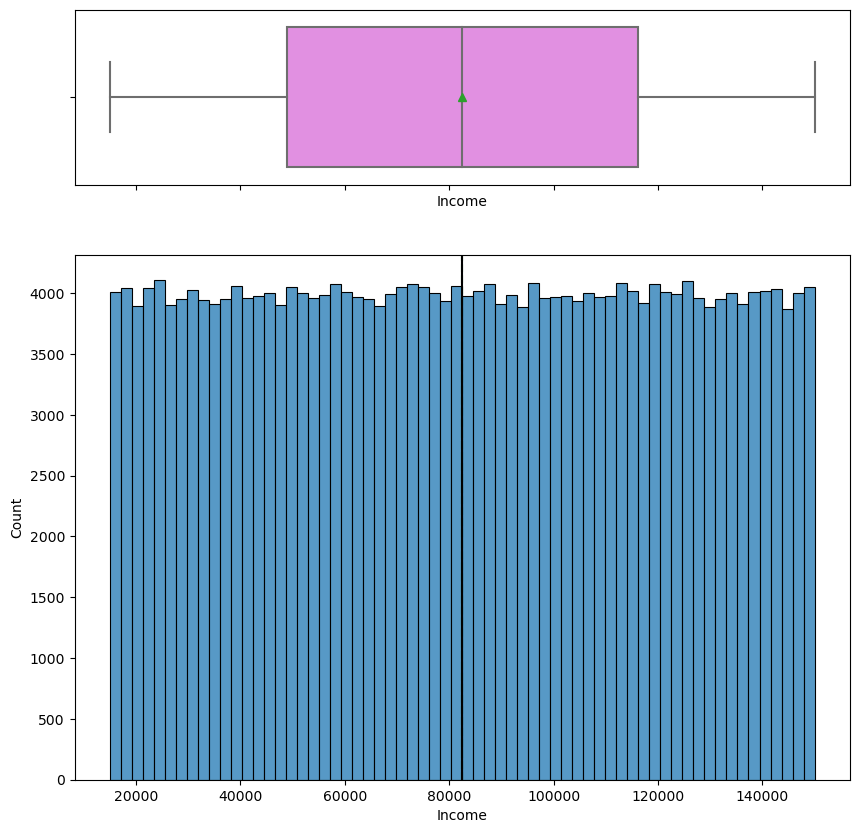
This graph shows that loan taken for business purpose have been defaulted the most. But as per our analysis, educational loan also has the second highest percentage in terms of default. So from this what we can infer is that people who passout from school and take loan for further studies or to start a business of their own have defaulted the maximum based on the above graphs.

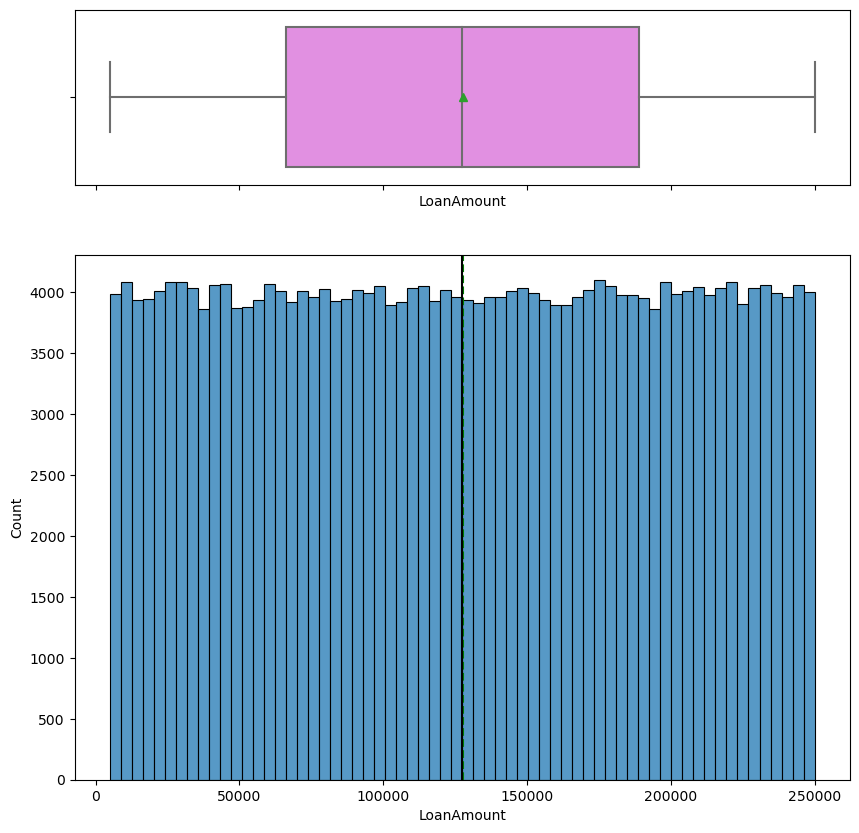
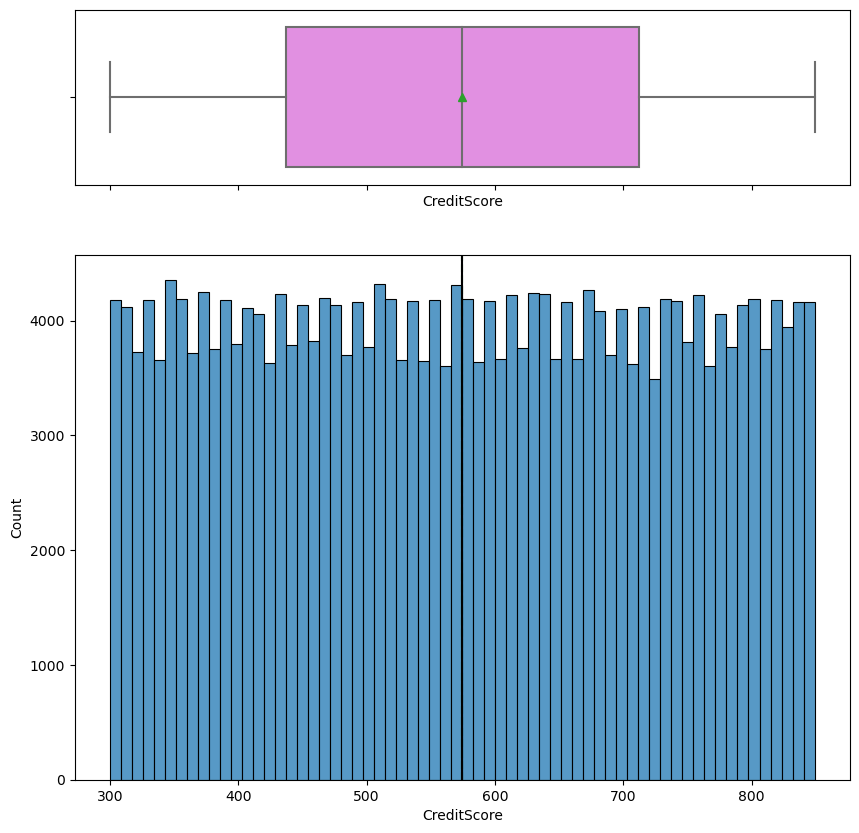


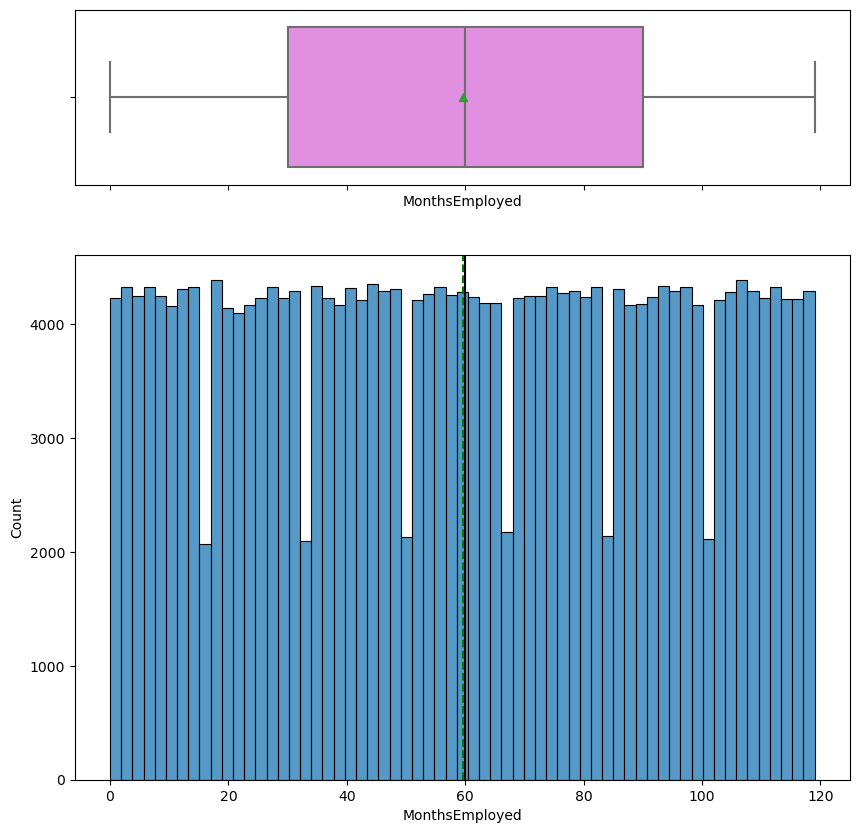
The chart above shows that people whitout any co signer are more likely to defaut a loan.



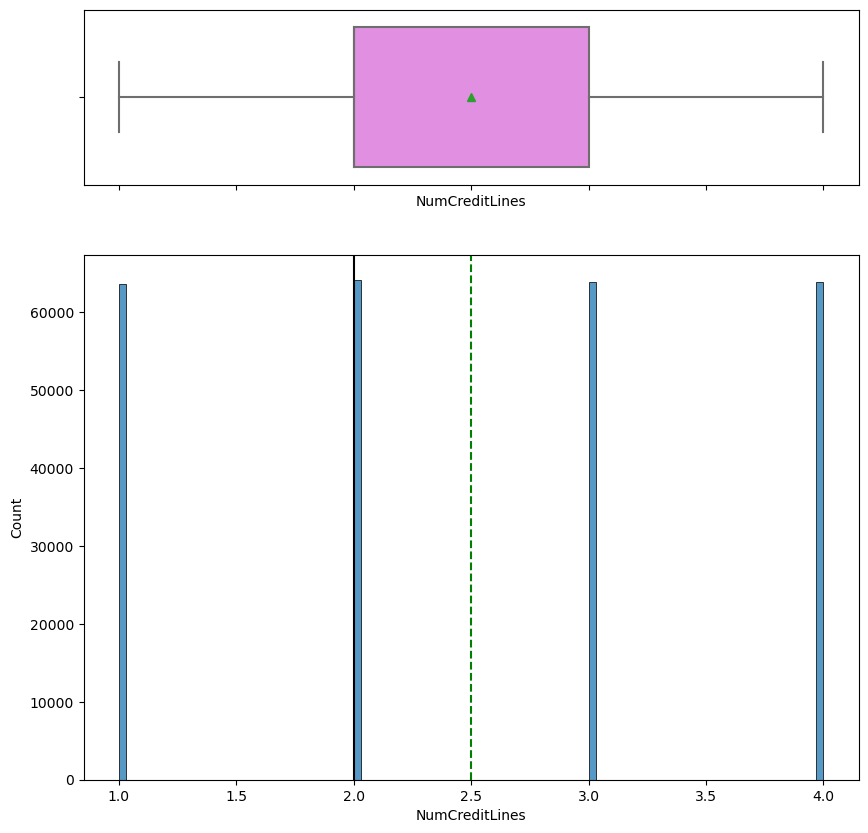
From the above plot we can see that the age variable is equally distributed as inferred by us in the data description section. The following graphs also show the same.



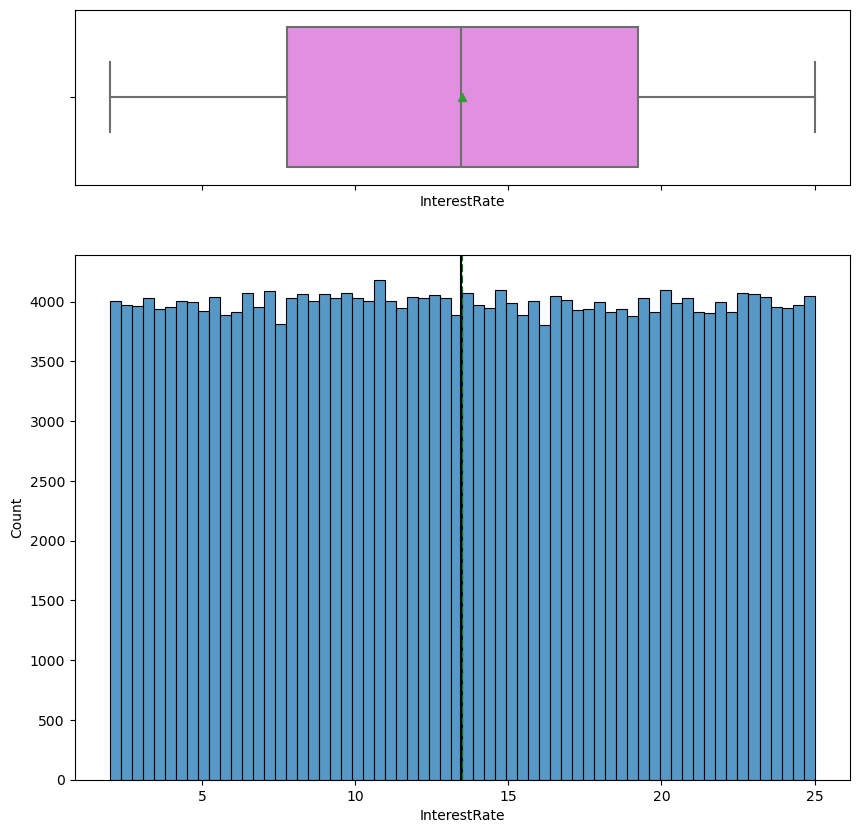
 

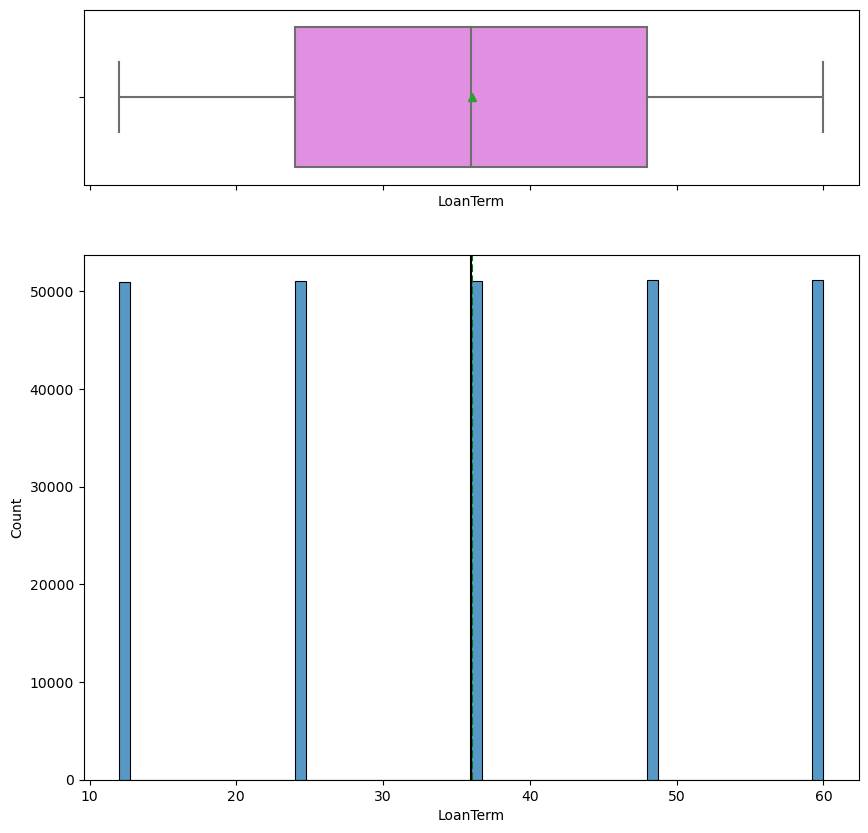


The above graph shows that there are many people with no work experience.

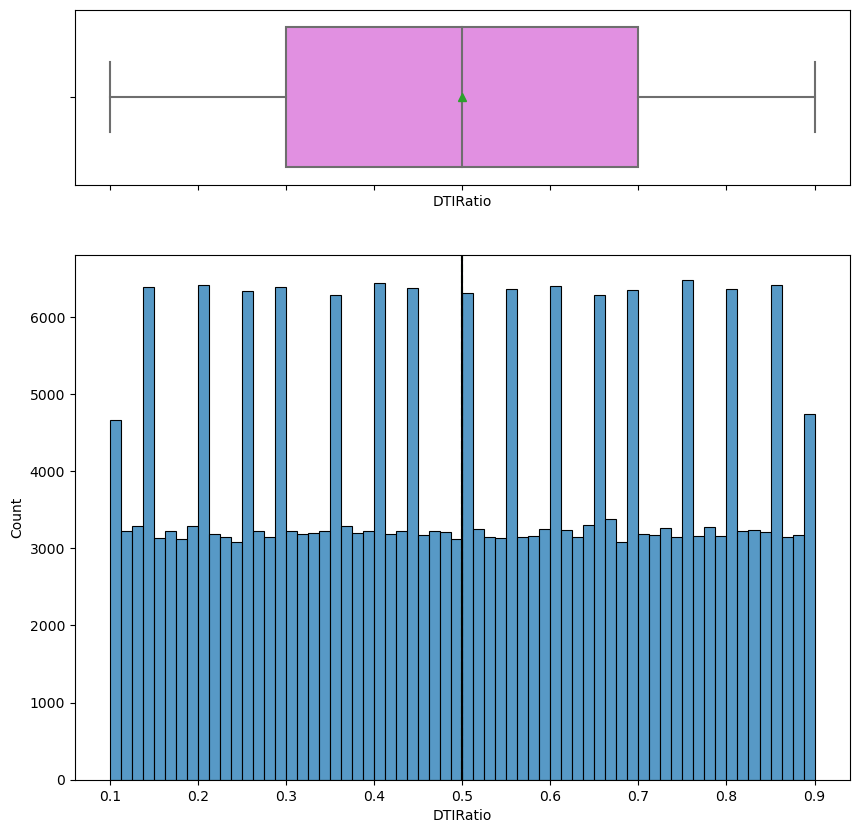


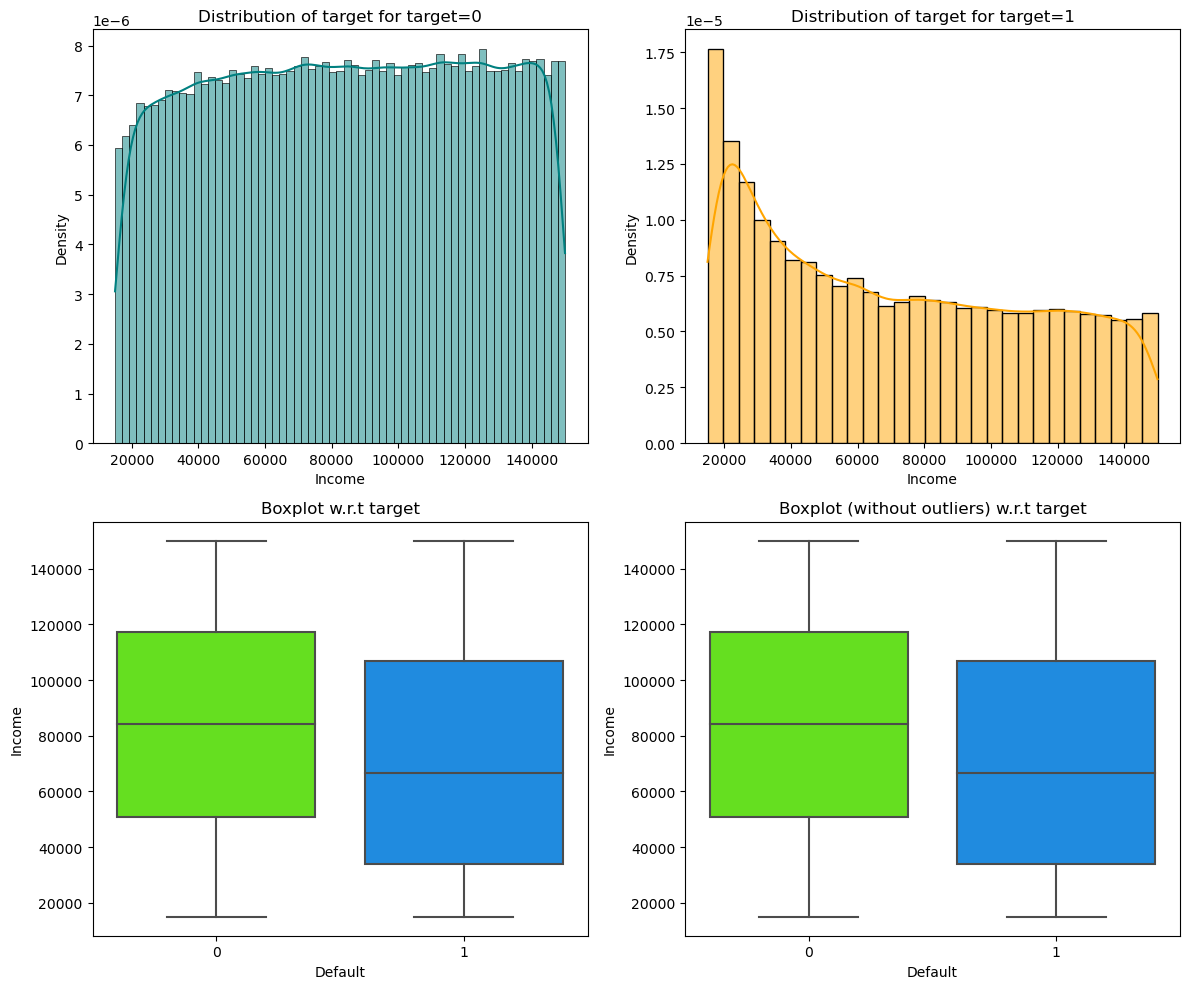
The number of credit lines that a customer has are 1 or 2 or 3 or 4.



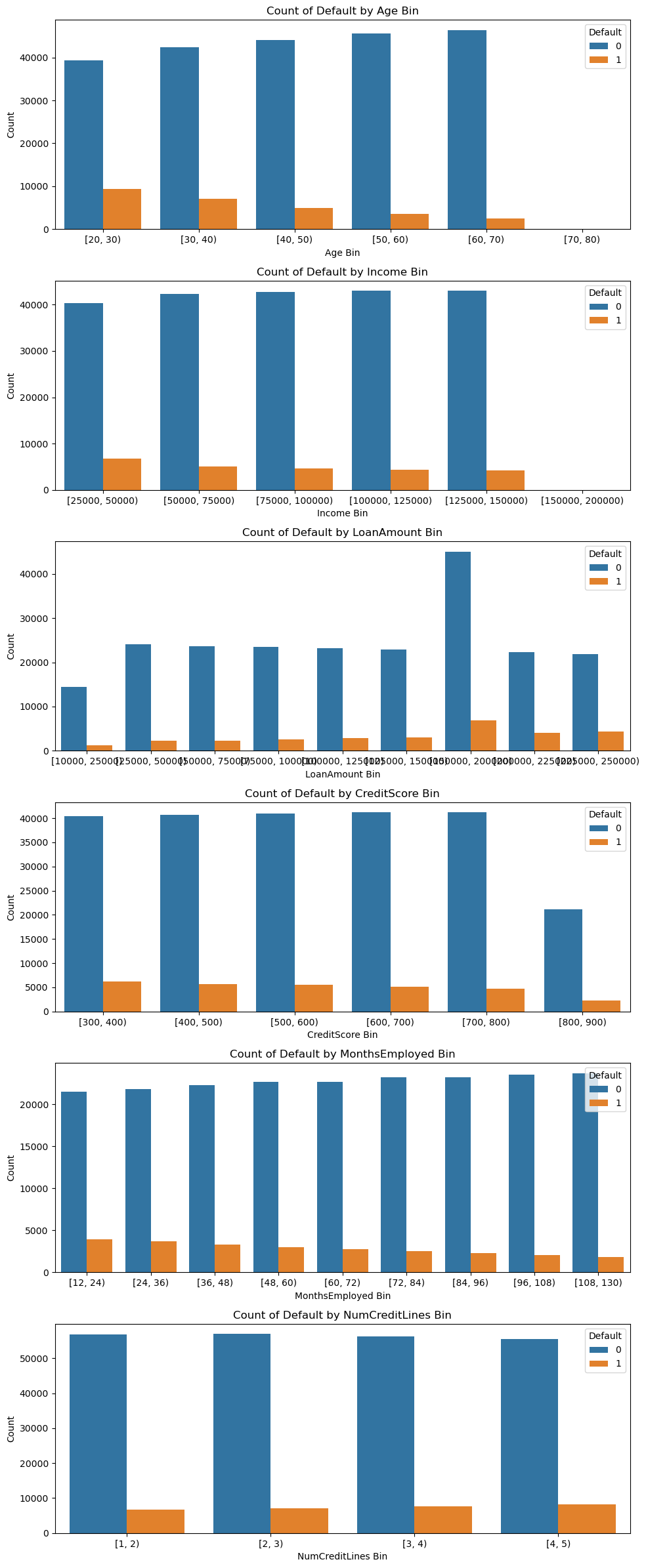


The loan terms are 12,24,36,48 and 60 months.



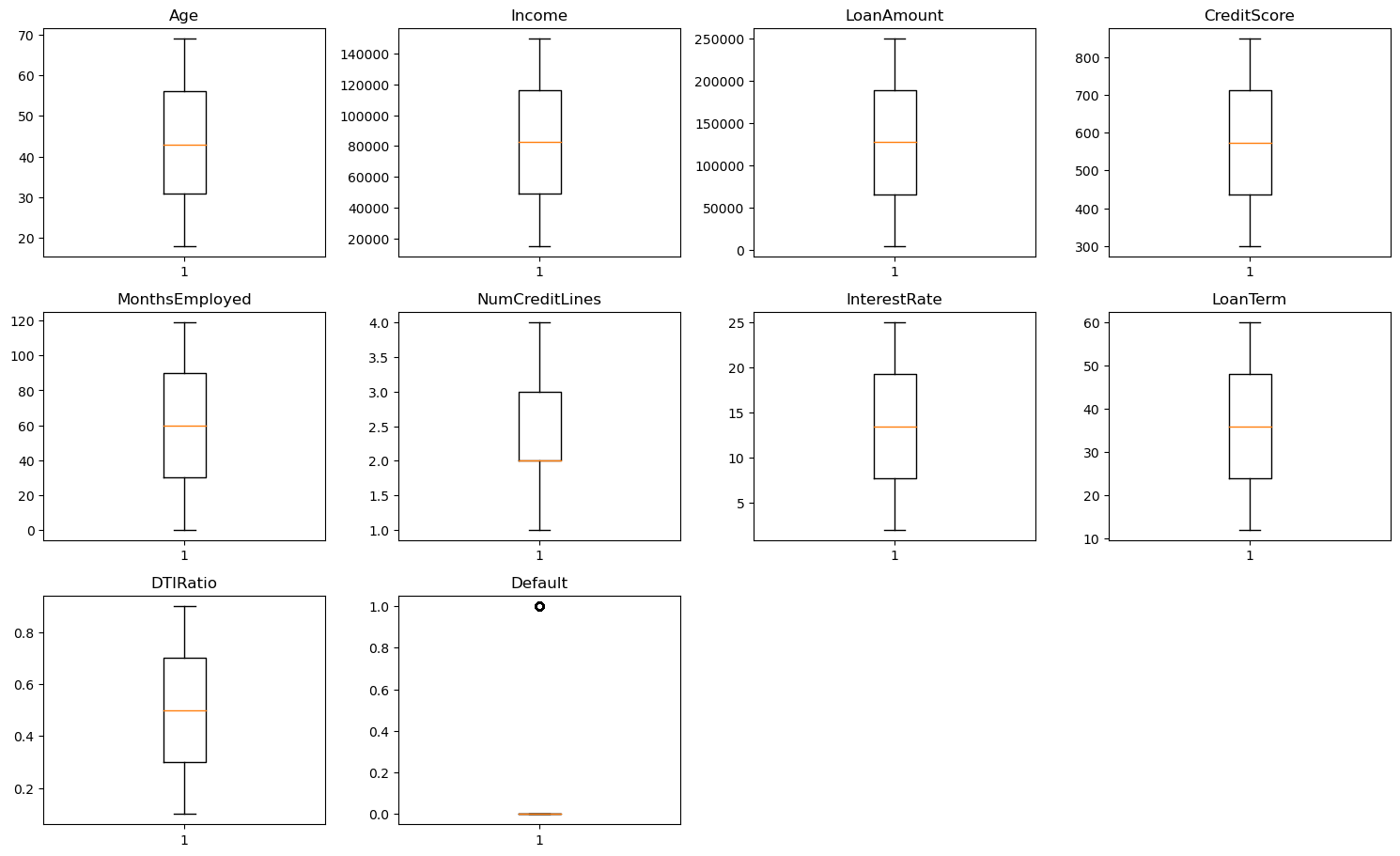


The above graph shows the relation between income and default. As we can infer that as income rises, the chances of defaulting goes down.

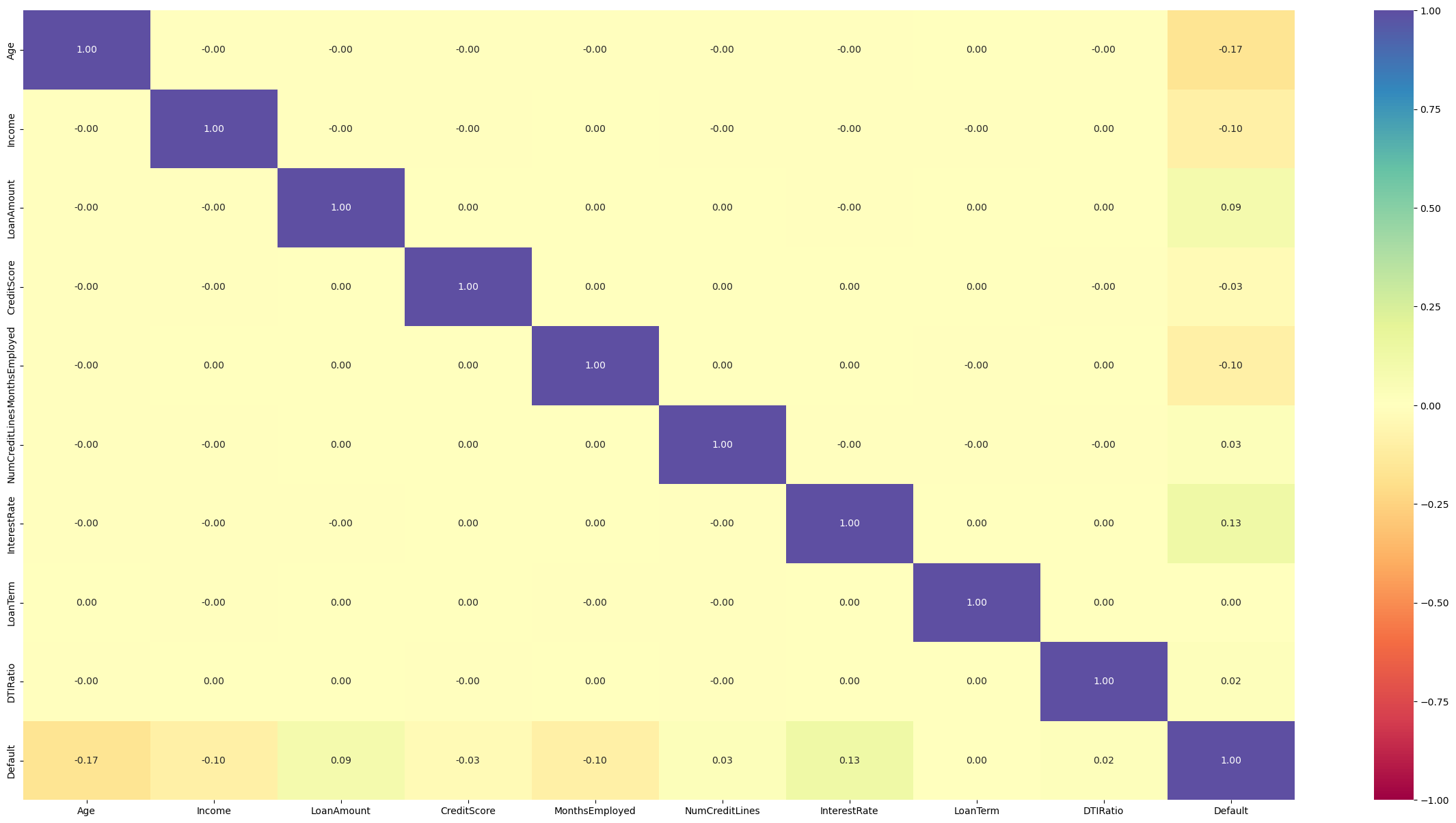


From the above graphs we can infer the following

* The more the age of the customer, the lesser is the chance of him/her defaulting. The most of the defaults in loan are done by people within the range of 18-30.
* As discussed earlier, the lesser the income, the more chances of the customer defaulting the loan.
* The more the loan amount taken by the customer, the greater are the chances of the customer defaulting.
* The lesser the credit score, more are the chances of the customers defaulting. Lower credit scores can also represent whether the customer has defaulted a loan earlier or has made late repayments.
* Also as the customer has more work experience the lesser is the chance of him/her defaulting.
* And as the number of credit lines that a customer has increases, the more are the chances of him/her defaulting.



From this graph we can see that there are no outliers and hence further data preprocessing is not needed.



Since we are dealing with a classification problem, the heat map shows us that there is no presence of multicollinearity in the daraset.