Bank Data set:
For the data wrangling ,Bank data set was used.
The columns in the dataset are continuous variable and of data type integer except for CCAvg Which is of float data type.

1>What types of cleaning method are used?

At the initial stage we remove the useless column ,ID, ,by using the code
df.drop('ID',axis=1,inplace=True)
The column ID doesnt serve any purpose for the observation .

The null values in the dataset was checked,using the code
df.isnull().sum()
Output:

```
Age                 0
Experience          0
Income              0
ZIP_Code            0
Family_members      0
CCAvg               0
Education           0
Mortgage            0
Personal_Loan       0
Securities_Account  0
CD_Account          0
Online              0
CreditCard          0
dtype: int64
```

It is observed that no missing values were found in the dataset

Any values that were duplicated in the dataset was then searched by using the code,
dupes = df.duplicated()
sum(dupes)
Output:0

It is observed that no such data was duplicated in the dataset .

2.Are there any missing values in the data?

df.isnull().sum()
Output:

```
Age                 0
Experience          0
Income              0
ZIP_Code            0
Family_members      0
CCAvg               0
Education           0
Mortgage            0
Personal_Loan       0
Securities_Account  0
CD_Account          0
Online              0
CreditCard          0
dtype: int64
```

All the columns did not have any missing values in the dataset.

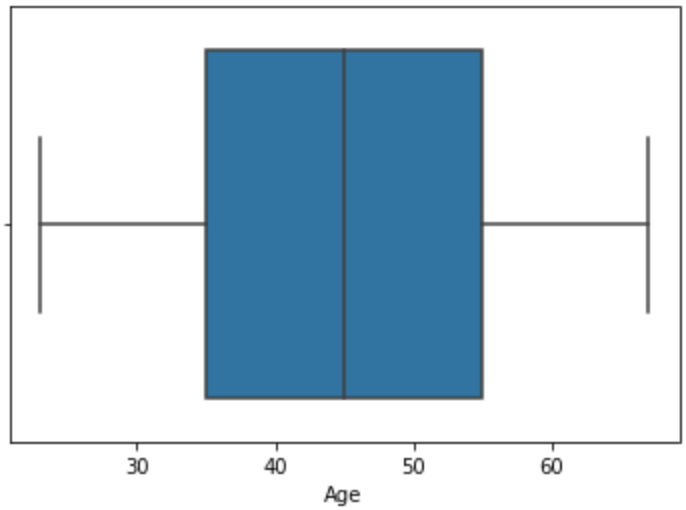
3.Check for the outliers,if any.How to treat the outliers?

For the dataset ,the code

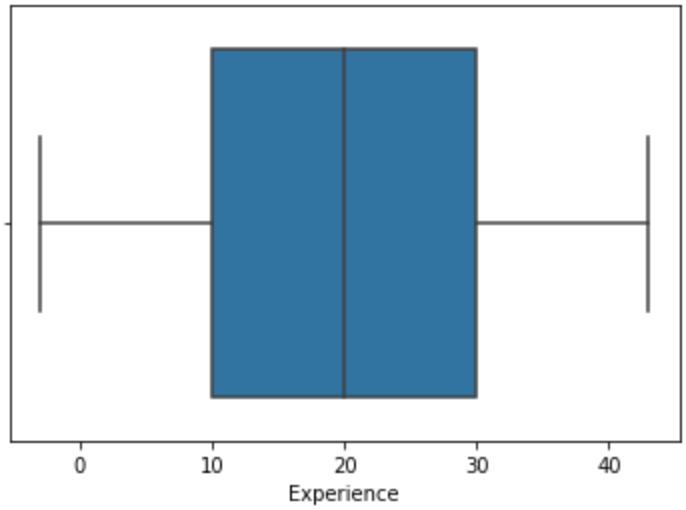
```
import seaborn as sns
sns.boxplot(x=df['Age'])
sns.boxplot(x=df['Experience'])
sns.boxplot(x=df['Income'])
sns.boxplot(x=df['ZIP_Code'])
sns.boxplot(x=df['Family_members'])
sns.boxplot(x=df['CCAvg'])
sns.boxplot(x=df['Education'])
sns.boxplot(x=df['Mortgage'])
sns.boxplot(x=df['Personal_Loan'])
sns.boxplot(x=df['Securities_Account'])
sns.boxplot(x=df['CD_Account'])
sns.boxplot(x=df['Online'])
sns.boxplot(x=df['CreditCard'])
```

Was used to find the outliers.

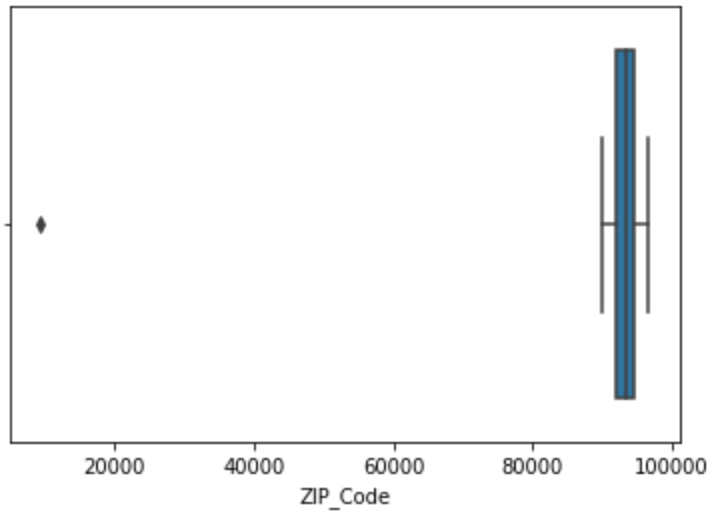
<matplotlib.axes._subplots.AxesSubplot at 0x1edefce0708>

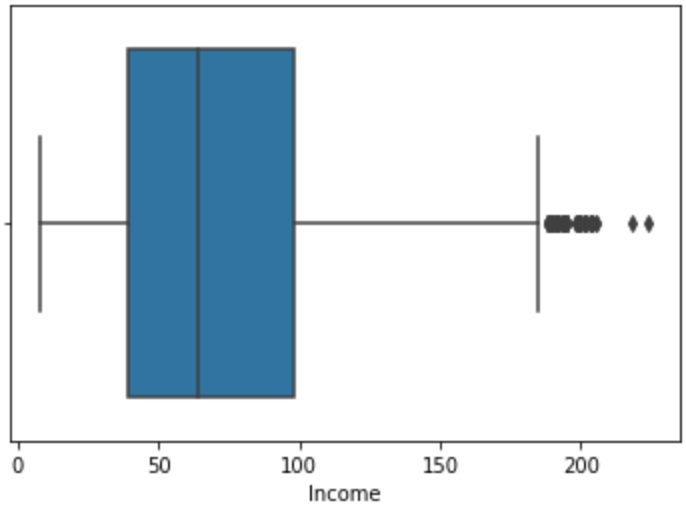



matplotlib.axes._subplots.AxesSubplot at 0x1edefd32448>

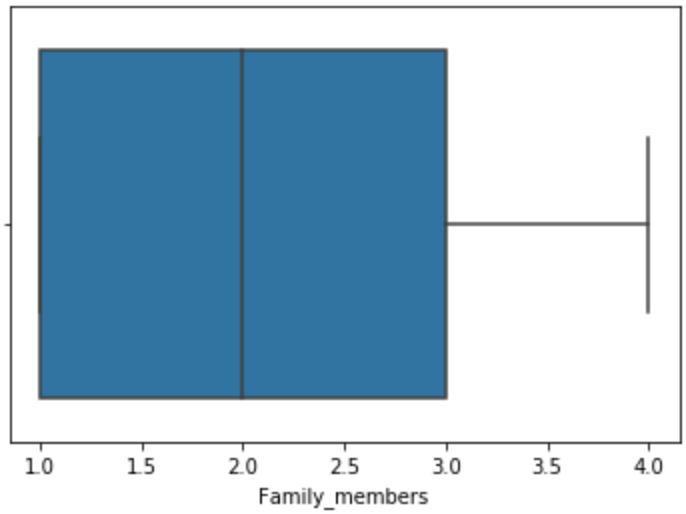



matplotlib.axes._subplots.AxesSubplot at 0x1edefd323c8>
matplotlib.axes._subplots.AxesSubplot at 0x1edefded1c8>

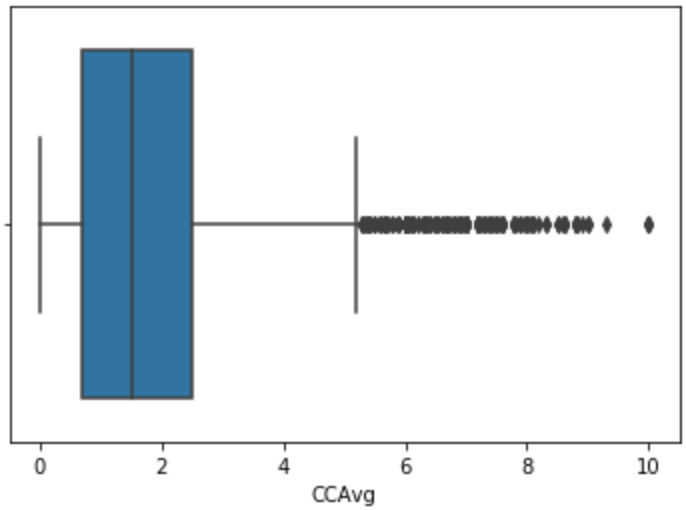


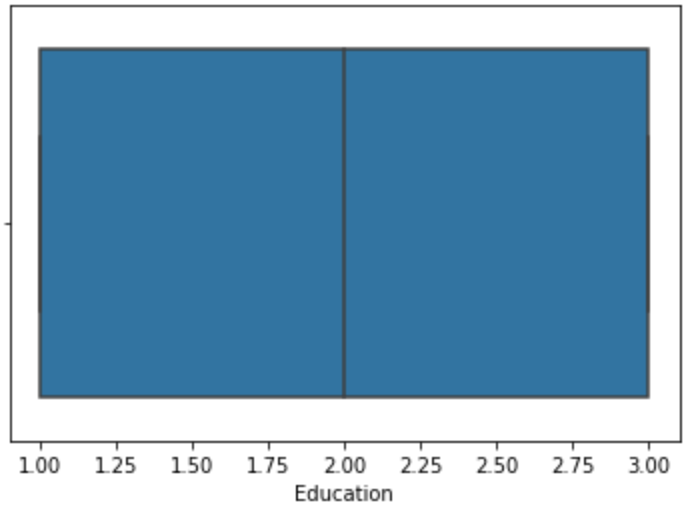

<matplotlib.axes._subplots.AxesSubplot at 0x1edefe55648>



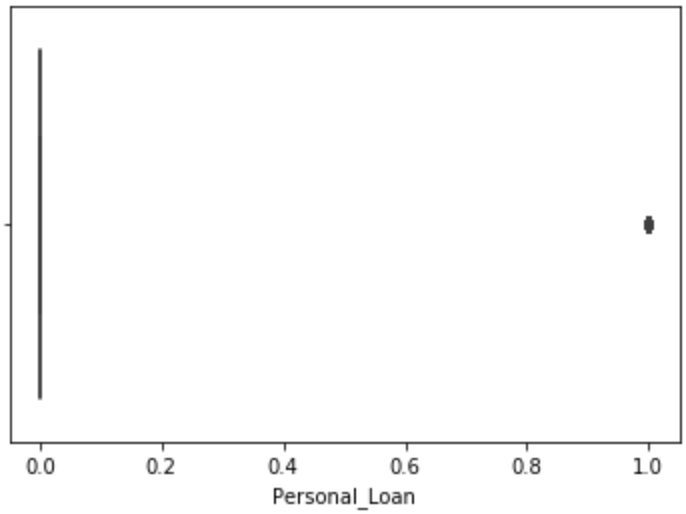
matplotlib.axes._subplots.AxesSubplot at 0x1edf1b98508>

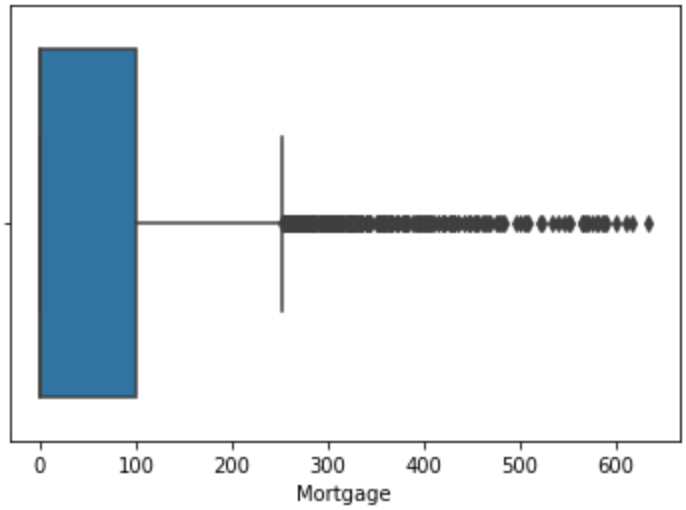



tplotlib.axes._subplots.AxesSubplot at 0x1edf1bed908>

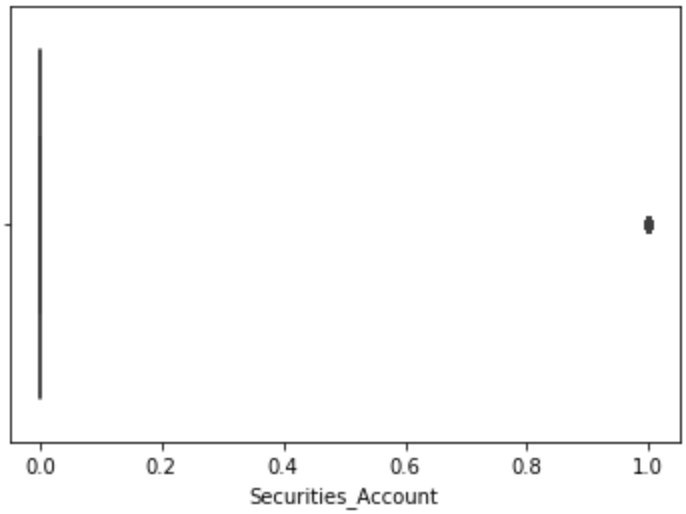



tplotlib.axes._subplots.AxesSubplot at 0x1edf1c6a108>
atplotlib.axes._subplots.AxesSubplot at 0x1edf1cd4348>

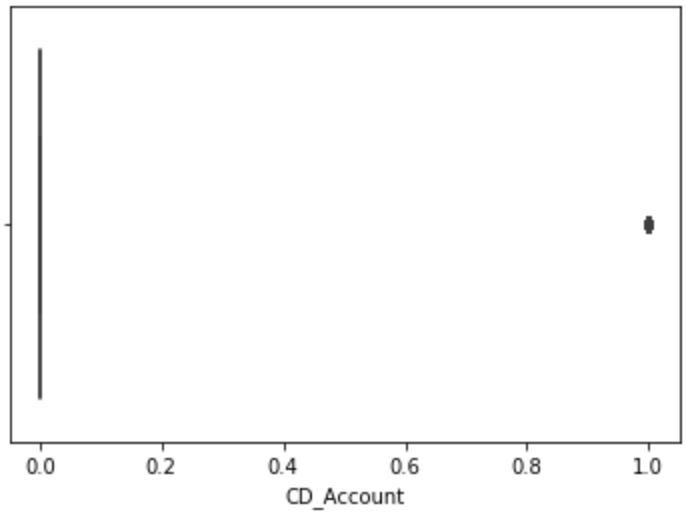


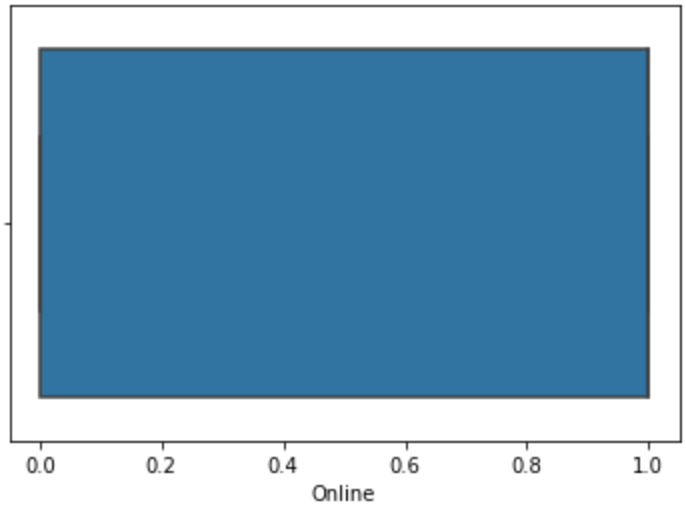

matplotlib.axes._subplots.AxesSubplot at 0x1edf1d187c8>



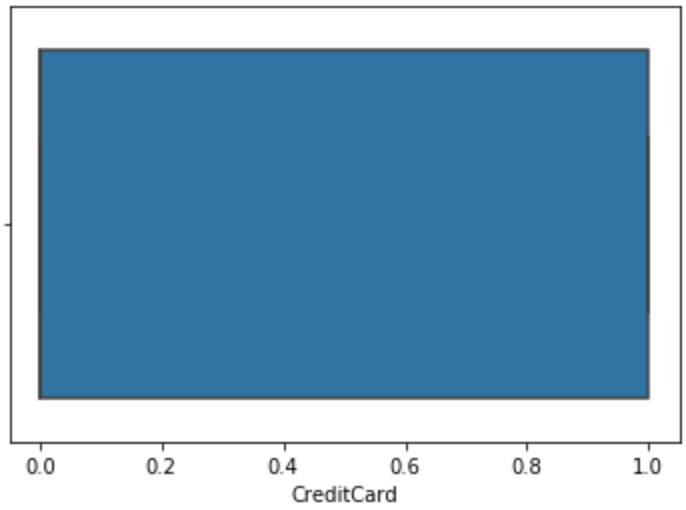
matplotlib.axes._subplots.AxesSubplot at 0x1edf1da7548>




<matplotlib.axes._subplots.AxesSubplot at 0x1edef4b13c8>




matplotlib.axes._subplots.AxesSubplot at 0x1edef56b7c8>




It is observed that Income,Zip code,ccavg,mortgage,personal_loan,securities_account and cd_account had outliers.
Age also had outliers,however very insignificant compared to the columns with outliers.
Experience,Family members,credit card and online did not have any outliers.

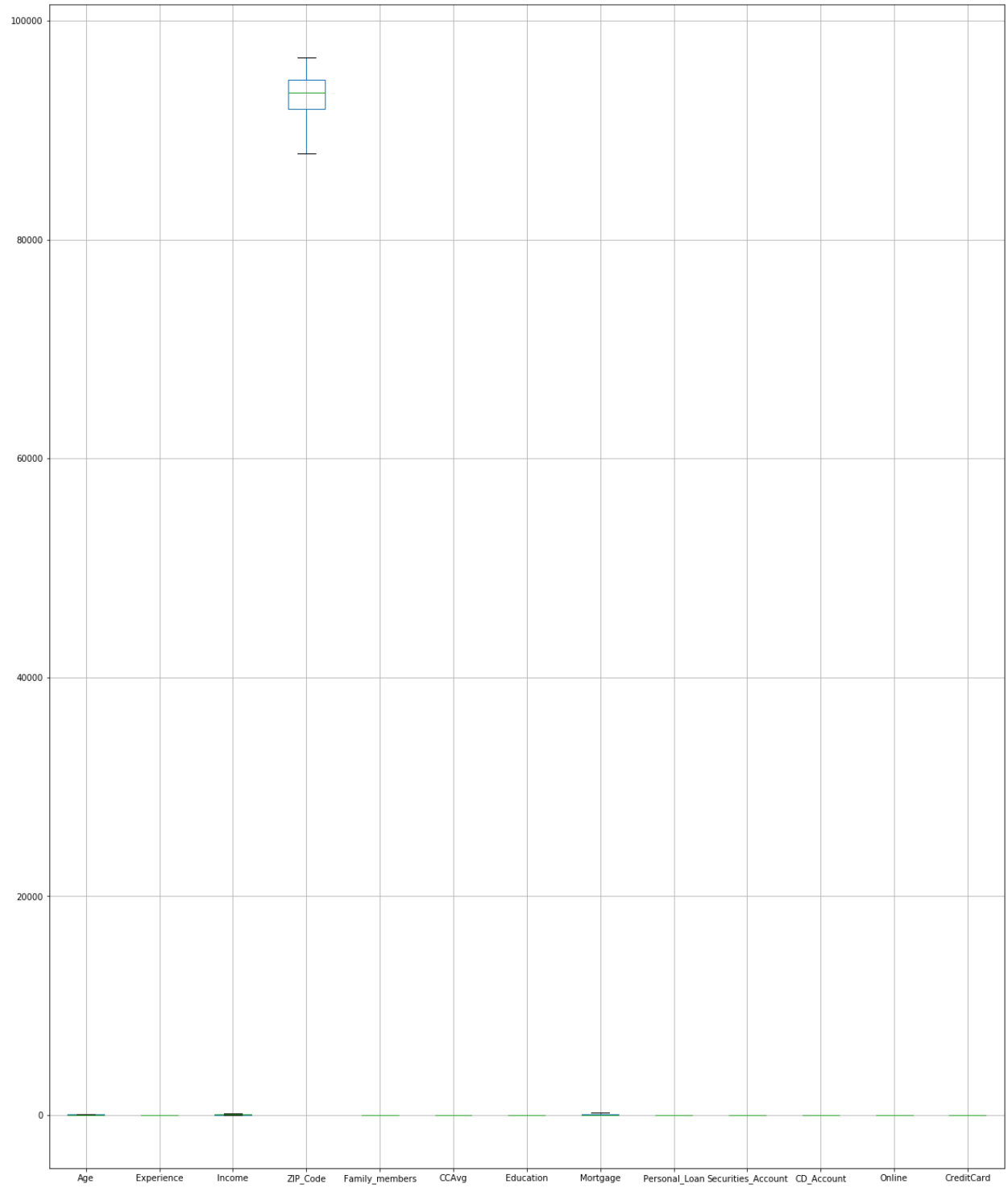
Treating the outliers.

We check for the interquartile ranges for the outliers,25% and 75% level.
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)
IQR = Q3 - Q1
print(IQR)
Output:

| | |
|---|---|
| Age | 20.0 |
| Experience | 20.0 |
| Income | 59.0 |
| ZIP_Code | 2697.0 |
| Family_members | 2.0 |
| CCAvg | 1.8 |
| Education | 2.0 |
| Mortgage | 101.0 |
| Personal_Loan | 0.0 |
| Securities_Account | 0.0 |
| CD_Account | 0.0 |
| Online | 1.0 |
| CreditCard | 1.0 |

dtype: float64
The interquartile ranges of the outliers of dataset is found.

After removing the outliers using the interquartile range :
Output:
<matplotlib.axes._subplots.AxesSubplot at 0x1edef88b408>




It is observed that most of the outliers were removed.
Ccavg,personal_loan,securities_account,cd_account ,zip_code had zero outliers.

Income,age,mortage had outliers but very insignificant.