Rocket Loans is an online personal lending company headquartered in the heart of downtown Detroit, Michigan. Rocket Loans is the sister company of Quicken Loans, an online lender based in Detroit and founded in 1985. Rocket Loans offers fast funding and competitive annual percentage rates (APRs). The convenient loan experience, however, has to be balanced against the fact that the company does charge an origination fee. Before deciding on a personal loan, you’ll want to do your own research to determine whether Rocket Loans is the best fit for your borrowing needs.

The company thereby needs a model developed which will help the agents to visit the right customer looking at the prediction that the model would make depending on the given data fields.

The firm also looks forward to some key analytical points that will help it/its agents to drive the business in the profitable direction.

Problem Statement: To predict Loan Approval (1 or 0) based on applicant's details.

|  |  |
| --- | --- |
| Sex | Gender(Male/Female) |
| Age | Age |
| Married | Marital status |
| No. of People in the Family Qualification | No. of People in the Family Qualification |
| Self\_Employed | If the person applying for loan is self employed |
| Loan\_Bearer\_Income | Income of the applicant |
| Loan\_Cobearer\_Income | Income of guarantor |
| Amount Disbursed | Amount provided |
| Loan\_Tenure | the amount of time taken to repay the loan. |
| Credit\_Score | Credit score |
| Location\_type | location(region) |
| Loan\_Status | If loan was provided or not |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Sex | Age | Married | No. of People in the Family | Qualification | Self\_Employed | Loan\_Bearer\_Income | Loan\_Cobearer\_Income | Amount Disbursed | Loan\_Tenure | Credit\_Score | Location\_type | Loan\_Status |
| 0 | Male | 33 | No | 0 | Graduate | No | 5849 | 0 | 108 | 360 | 1 | Urban | Y |
| 1 | Male | 34 | Yes | 1 | Graduate | No | 4583 | 1508 | 128 | 360 | 1 | Rural | N |
| 2 | Male | 31 | Yes | 0 | Graduate | Yes | 3000 | 0 | 66 | 360 | 1 | Urban | Y |
| 3 | Male | 33 | Yes | 0 | Not Graduate | No | 2583 | 2358 | 120 | 360 | 1 | Urban | Y |
| 4 | Male | 48 | No | 0 | Graduate | No | 6000 | 0 | 141 | 360 | 1 | Urban | Y |

Data Sample:

**Exploratory Data Analysis – EDA**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Column | Non-Null Count | Dtype |
|  |  |  |  |
| 1 | Sex | 601 non-null | object |
| 2 | Age | 614 non-null | int64 |
| 3 | Married | 611 non-null | object |
| 4 | No. of People in the Family | 599 non-null | object |
| 5 | Qualification | 614 non-null | object |
| 6 | Self\_Employed | 582 non-null | object |
| 7 | Loan\_Bearer\_Income | 614 non-null | int64 |
| 8 | Loan\_Cobearer\_Income | 614 non-null | float64 |
| 9 | Amount Disbursed | 593 non-null | float64 |
| 10 | Loan\_Tenure | 600 non-null | float64 |
| 11 | Credit\_Score | 564 non-null | float64 |
| 12 | Location\_type | 614 non-null | object |
| 13 | Loan\_Status | 614 non-null | object |

**Analysis:**

* We have 614 records
* We have total of 13 features.
* We have null values.
* Out of 13 features 9 are categorical in nature and remaining 4 are continuous in nature

**Features with null values:**

|  |  |
| --- | --- |
| Features | Count |
| Sex | 13 |
| Age | 0 |
| Married | 3 |
| No. of People in the Family | 15 |
| Qualification | 0 |
| Self\_Employed | 32 |
| Loan\_Bearer\_Income | 0 |
| Loan\_Cobearer\_Income | 0 |
| Amount Disbursed | 21 |
| Loan\_Tenure | 14 |
| Credit\_Score | 50 |
| Location\_type | 0 |
| Loan\_Status | 0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Age | Loan\_Bearer\_Income | Loan\_Cobearer\_Income | Amount Disbursed | Loan\_Tenure | Credit\_Score |
| count | 614 | 614 | 614 | 593 | 600 | 564 |
| mean | 38.46906 | 5403.459 | 1621.246 | 146.347386 | 342 | 0.842199 |
| std | 8.124732 | 6109.042 | 2926.248 | 85.529555 | 65.12041 | 0.364878 |
| min | 25 | 150 | 0 | 9 | 12 | 0 |
| 25% | 32 | 2877.5 | 0 | 100 | 360 | 1 |
| 50% | 39 | 3812.5 | 1188.5 | 128 | 360 | 1 |
| 75% | 45 | 5795 | 2297.25 | 168 | 360 | 1 |
| max | 52 | 81000 | 41667 | 700 | 480 | 1 |

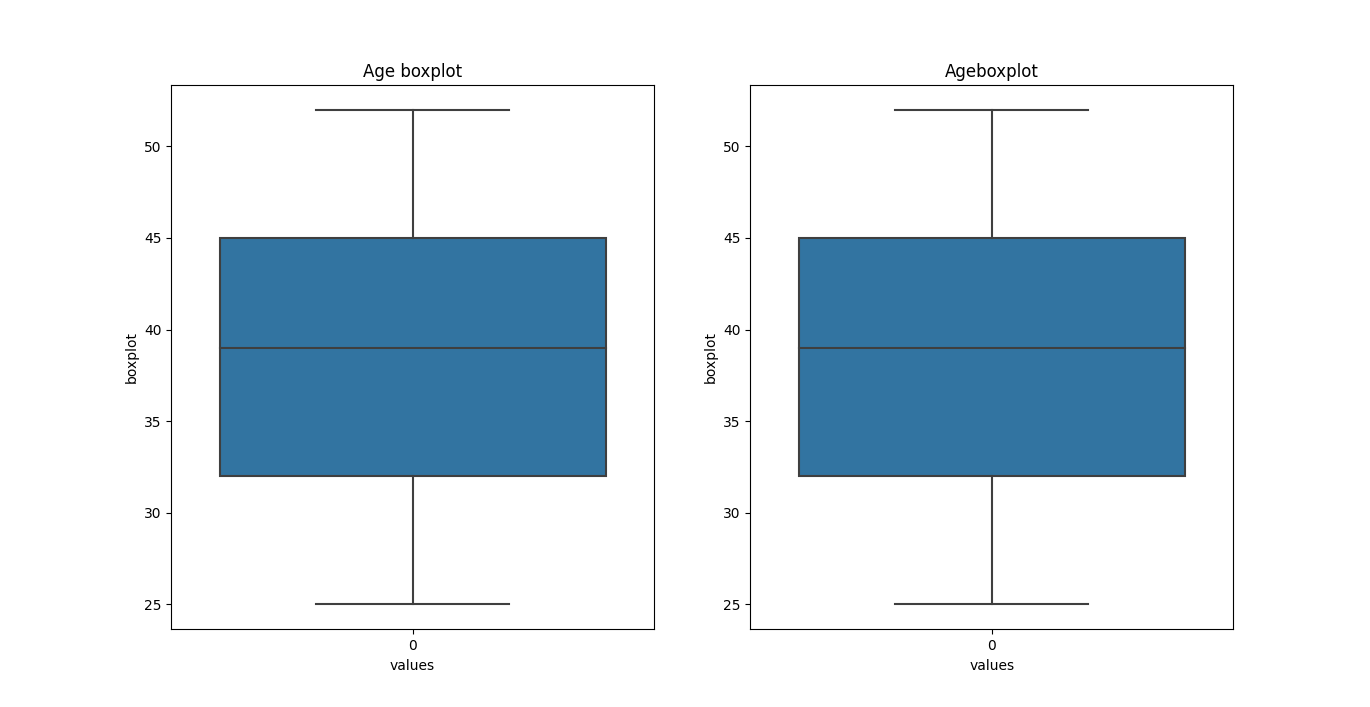
**Descriptive Statistics:**

**Analysis:**

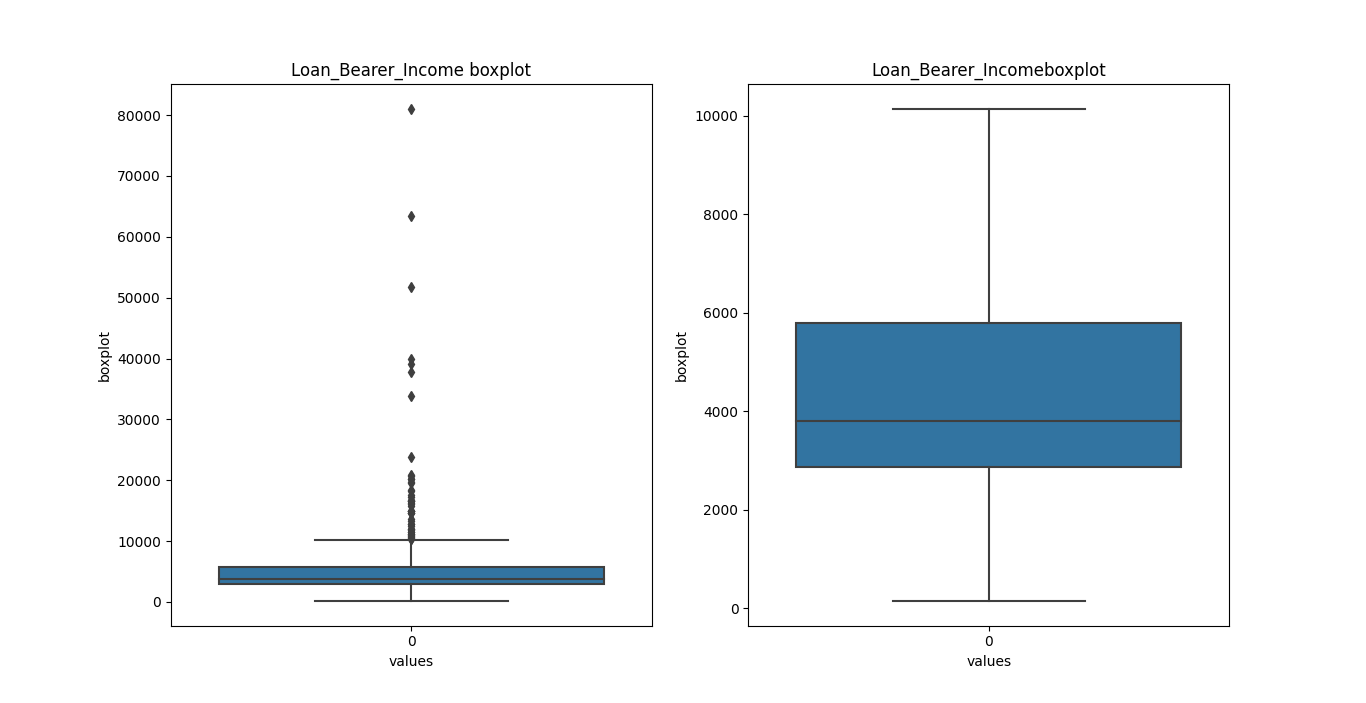
* Age may not have outliers as mean and median are almost same.
* For Loan bearer income there is no chance of outliers in the lower whisker region as min and Q1 are same.
* Loan Bearer Income, Loan Cobearer Income, Amount Disbursed, Loan Tenure have outliers.

**Outliers Treatment:**

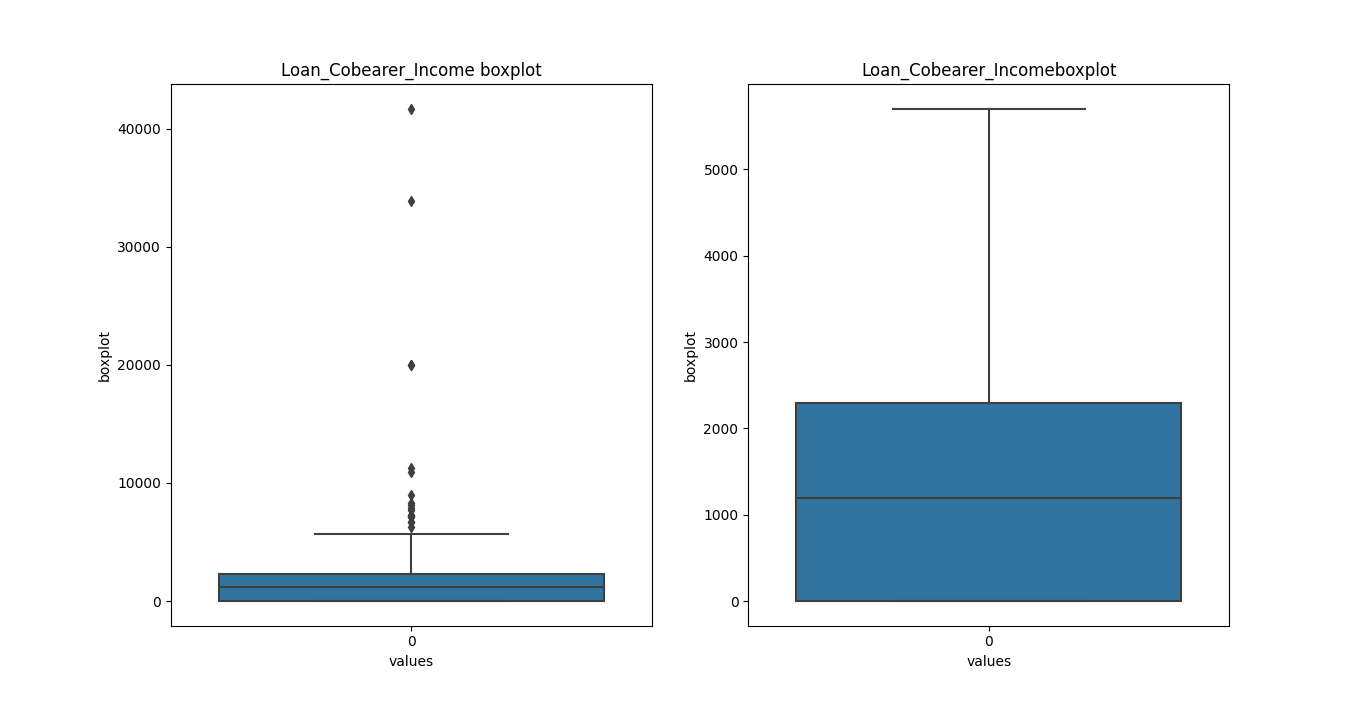
Age:

* Feature Age does not have any outliers

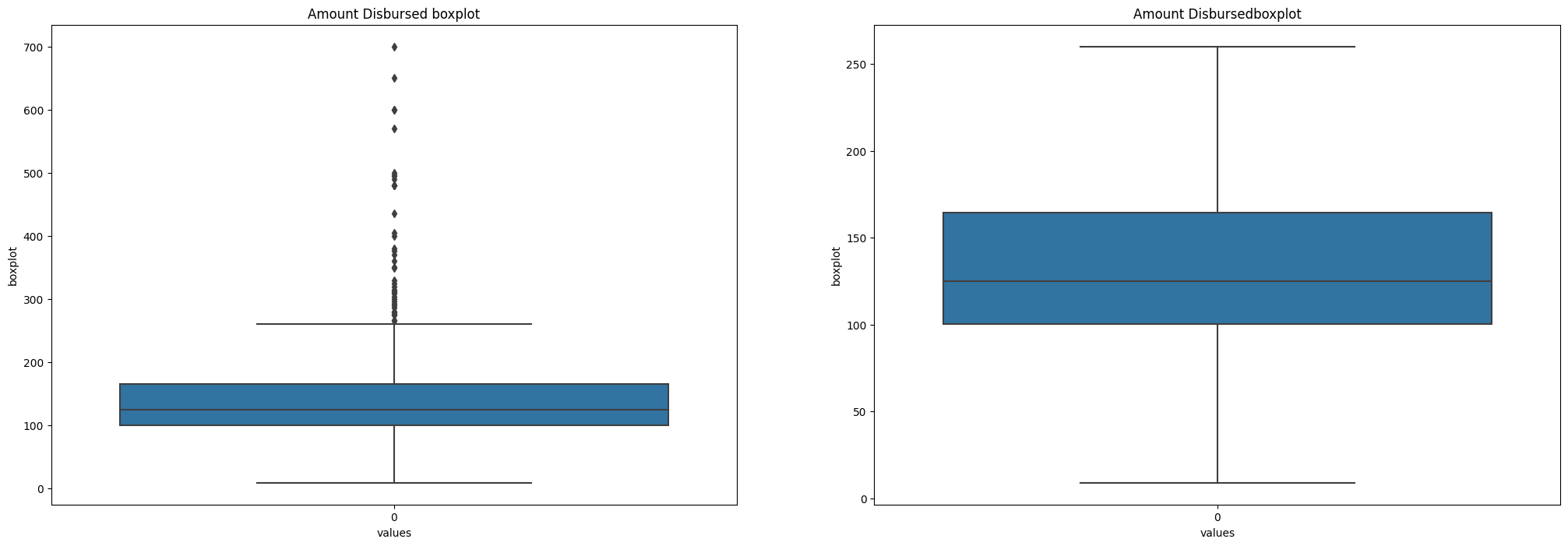
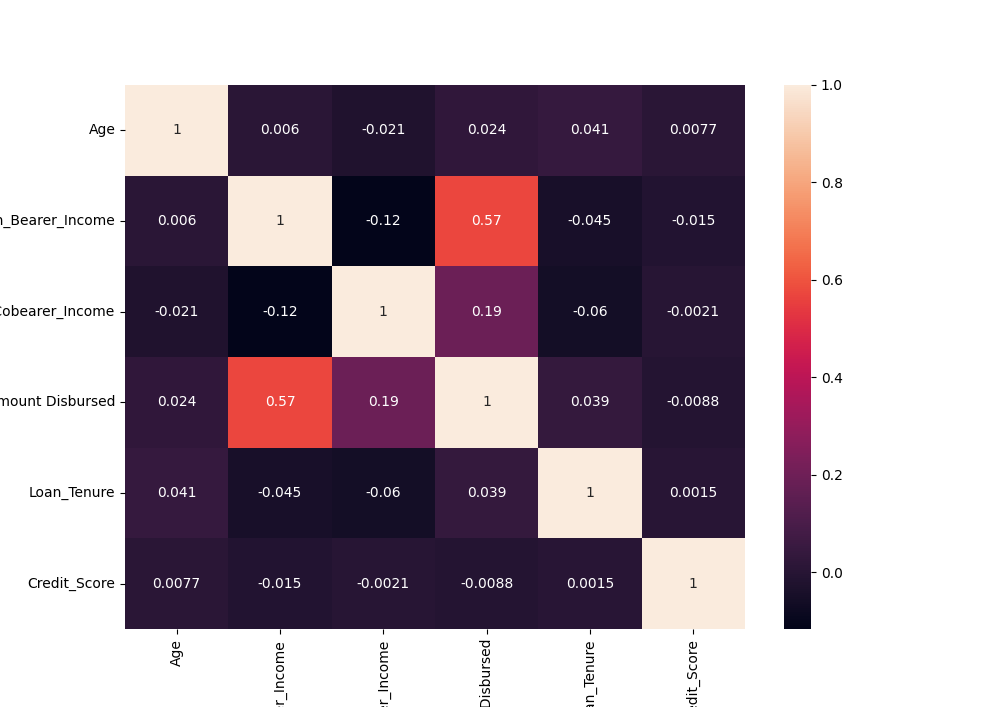
Loan bearer income:

* Feature Loan\_Bearer\_Income has outliers
* Total number of outliers in this Loan\_Bearer\_Income is: 50
* Outliers percentage in Loan\_Bearer\_Income is 8.0%

**Loan Cobearer Income**

* Feature Loan\_Cobearer\_Income has outliers
* Total number of outliers in this Loan\_Cobearer\_Income is: 18
* Outliers percentage in Loan\_Cobearer\_Income is 3.0%

**Amount Disbursed:**

* Feature Amount Disbursed has outliers
* Total number of outliers in this Amount Disbursed is: 41
* Outliers percentage in Amount Disbursed is 7.000000000000001%
* **Multi-collinearity-Check**
* **Stage-1: Correlation heatmap**

Analysis:

* Loan\_Bearer\_Income and amount disbursed have higher correlation.

All remaining features have collinearity lesser than 30% which indicates that there is less chance of multicollinearity.

|  |  |  |
| --- | --- | --- |
|  | VIF | Independent\_variables |
| 9 | 18.576546 | Loan\_Tenure |
| 1 | 17.329317 | Age |
| 8 | 14.657214 | Amount Disbursed |
| 6 | 9.64887 | Loan\_Bearer\_Income |
| 0 | 6.530021 | Sex |
| 10 | 6.456034 | Credit\_Score |
| 2 | 3.859877 | Married |
| 11 | 2.648211 | Location\_type |
| 7 | 2.513374 | Loan\_Cobearer\_Income |
| 3 | 1.827661 | No. of People in the Family |
| 4 | 1.324599 | Qualification |
| 5 | 1.21651 | Self\_Employed |

**Multicollinearity test**

* Stage 2: Variance Inflating Factor (VIF)

formula for VIF = 1/(1-R2)

1. Regress every independent variable with each other and find the R2 score

2. find out VIF using above formula

3. if VIF is more than 5 for any independent variable we can conclude that multi-collinearity exists.

**Analysis**

* Loan\_Tenure, Age, Amount Disbursed, Loan\_Bearer\_Income, Sex, Credit\_Score have VIF score >5 so we can conclude that multicollinearity exists.

**Correlation with target feature:**

|  |  |  |
| --- | --- | --- |
| Position | independent | variables |
| 10 | Credit\_Score | 0.540556 |
| 2 | Married | 0.091478 |
| 1 | Age | 0.061043 |
| 7 | Loan\_Cobearer\_Income | 0.033237 |
| 11 | Location\_type | 0.032112 |
| 0 | Sex | 0.017987 |
| 6 | Loan\_Bearer\_Income | 0.010462 |
| 3 | No. of People in the Family | 0.010118 |
| 5 | Self\_Employed | -0.0037 |
| 9 | Loan\_Tenure | -0.022549 |
| 8 | Amount Disbursed | -0.040674 |
| 4 | Qualification | -0.085884 |

**Analysis:**

* Credit score has higher correlation with target feature.

**Feature engineering**

Applying Principal component analysis and adding target feature

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | PC\_1 | PC\_2 | PC\_3 | PC\_4 | PC\_5 | PC\_6 | PC\_7 | PC\_8 | PC\_9 | PC\_10 |
| 0 | -0.99488 | 1.179673 | 0.244201 | -1.25488 | 0.065709 | 0.222998 | -0.1942 | 0.187052 | 1.094624 | -1.08727 |
| 1 | 0.806848 | -0.18679 | -0.63987 | 0.168663 | 0.12435 | -0.49363 | -1.08619 | 0.131014 | -0.71006 | -0.65546 |
| 2 | -0.50796 | 0.218361 | 1.029491 | -0.90948 | -0.00239 | -0.63204 | 1.254438 | -2.43862 | -0.21153 | -1.77354 |
| 3 | -0.34620 | -1.86347 | -0.09183 | -0.32284 | 0.266685 | -0.32262 | 0.110599 | -0.84507 | 1.907777 | 0.216322 |
| 4 | -0.70018 | 1.404674 | 0.09877 | -0.78663 | -1.3816 | 0.80057 | 0.321335 | 0.700613 | 1.081574 | -0.37488 |

Analysis:

* Total number of features were reduced to 10 from 12
* Above features are considered for model building and evaluation.
* Explained variance ratio after PCA is:

[0.27182784 0.44776951 0.61076485 0.73733075 0.86207842 0.97452865]

**Model building:**

**1. Train-test split –** Splitting the data into train and test sets

**2. Cross-validation -** Evaluating machine learning models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data.

**3. Hyperparameter tuning-** Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

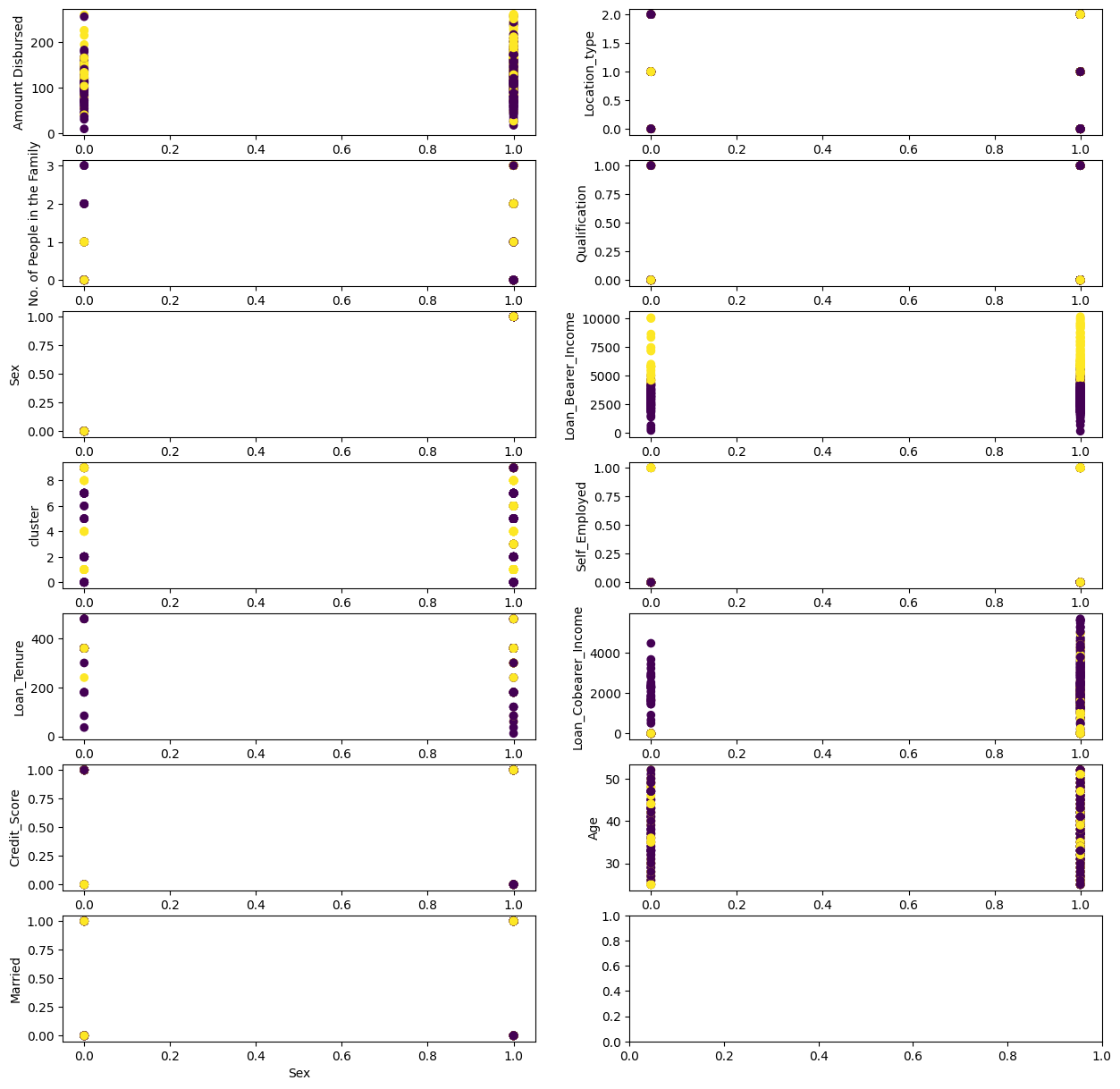
|  |
| --- |
| model: RF |
| Best\_params: {'max\_depth': 3, 'max\_features': 2, 'n\_estimators': 150} |
| model: XB |
| Best\_params: {'alpha': 10, 'eta': 1, 'gamma': 0, 'max\_depth': 3, 'reg\_lambda': 10} |
| model: GB |
| Best\_params: {'learning\_rate': 0.1, 'n\_estimators': 50} |
| model: KNN |
| Best\_params: {'algorithm': 'auto', 'weights': 'uniform'} |
| model: SVR |
| Best\_params: {'C': 0.1, 'gamma': 'scale'} |
| model: DTR |
| Best\_params: {'criterion': 'gini', 'max\_depth': 3, 'max\_features': 6} |
| model: AB |
| Best\_params: {'learning\_rate': 0.1, 'n\_estimators': 150} |

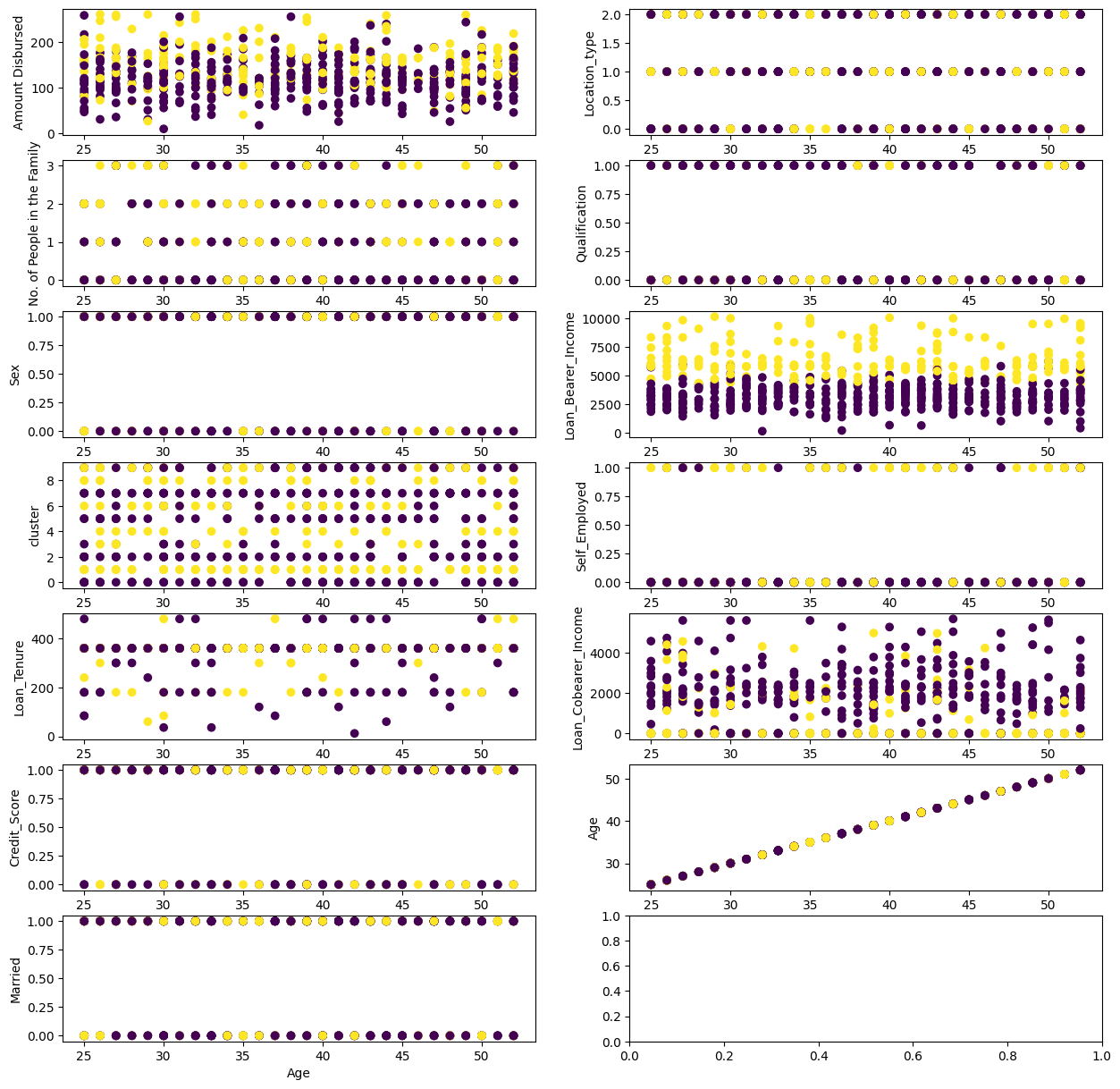
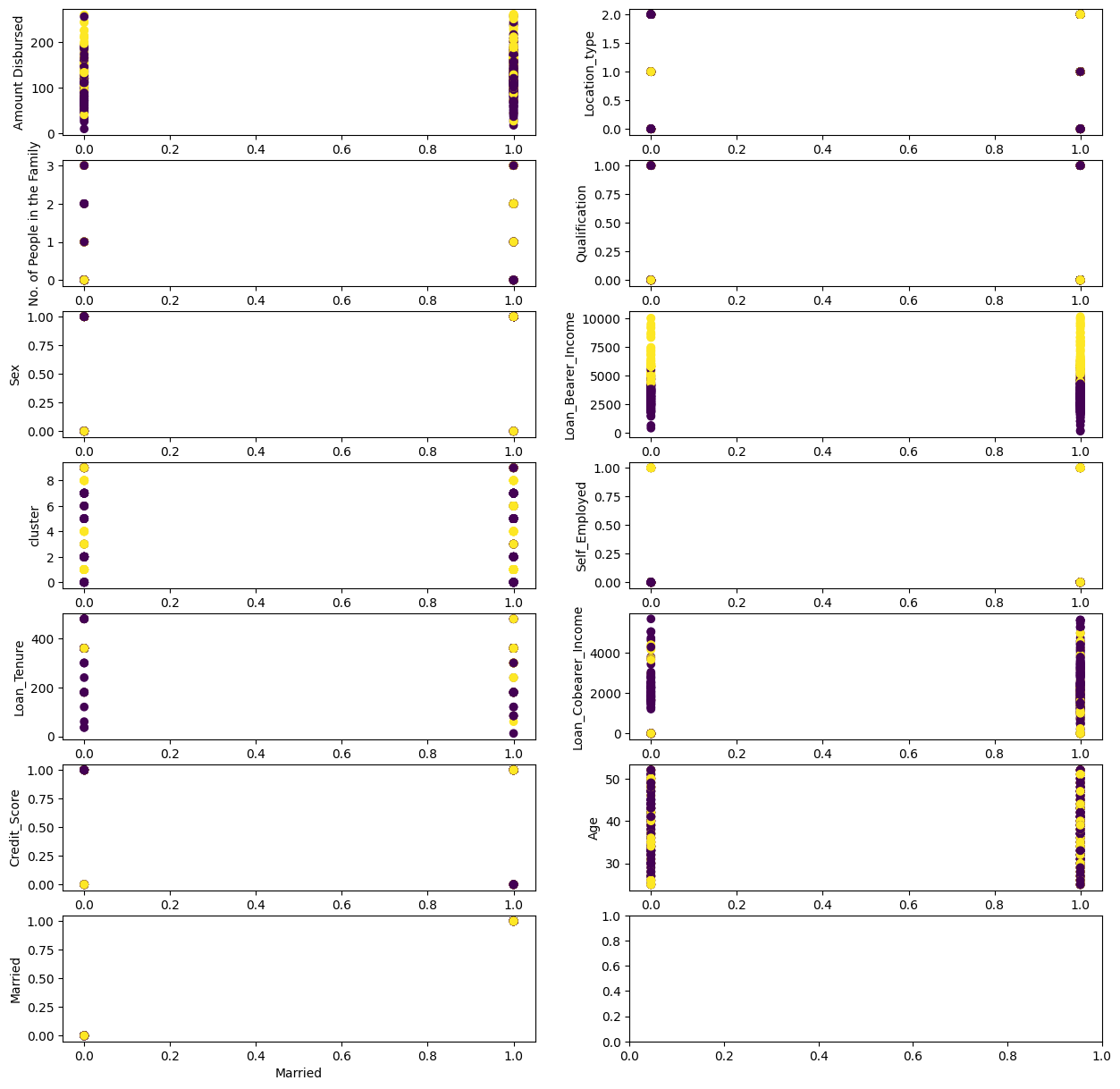
**Using clustering to check if it can help us improve accuracy**

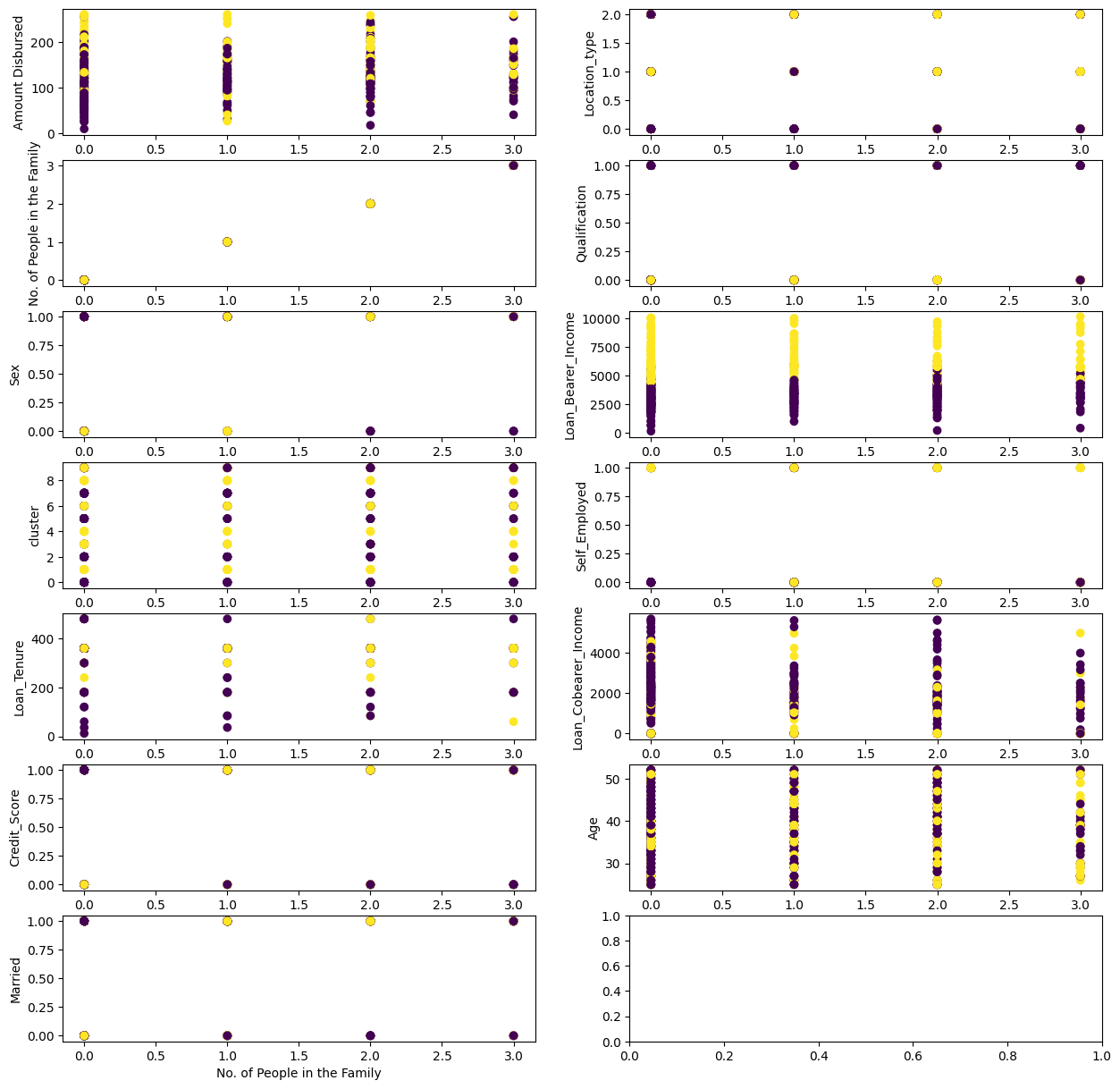
Checking features with target to see which one makes clear clusters

We are trying to use independent variables to see if there is any clusters being formed in the data so that we can use that as another independent variable in evaluation.

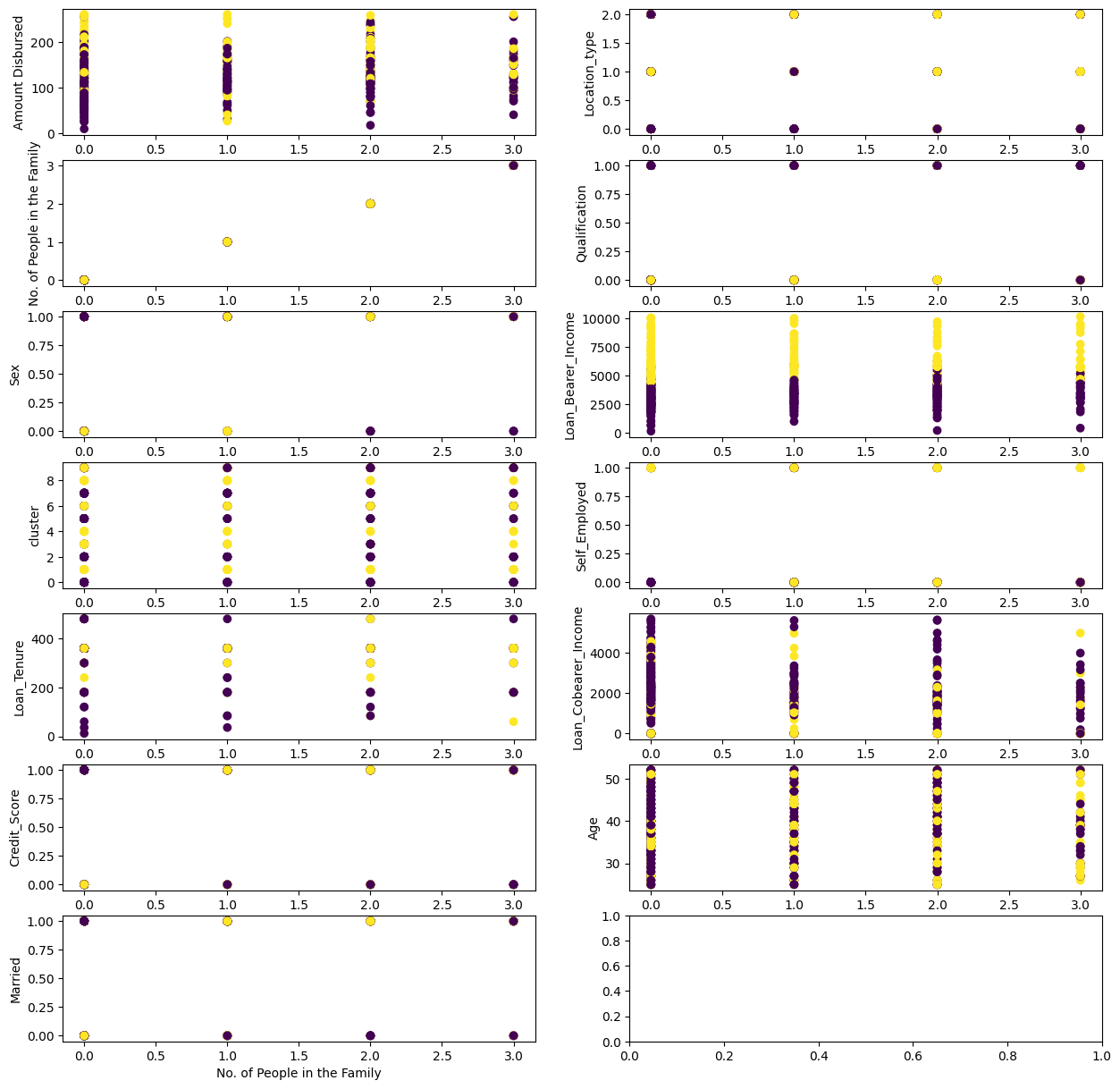
1. **Sex:**



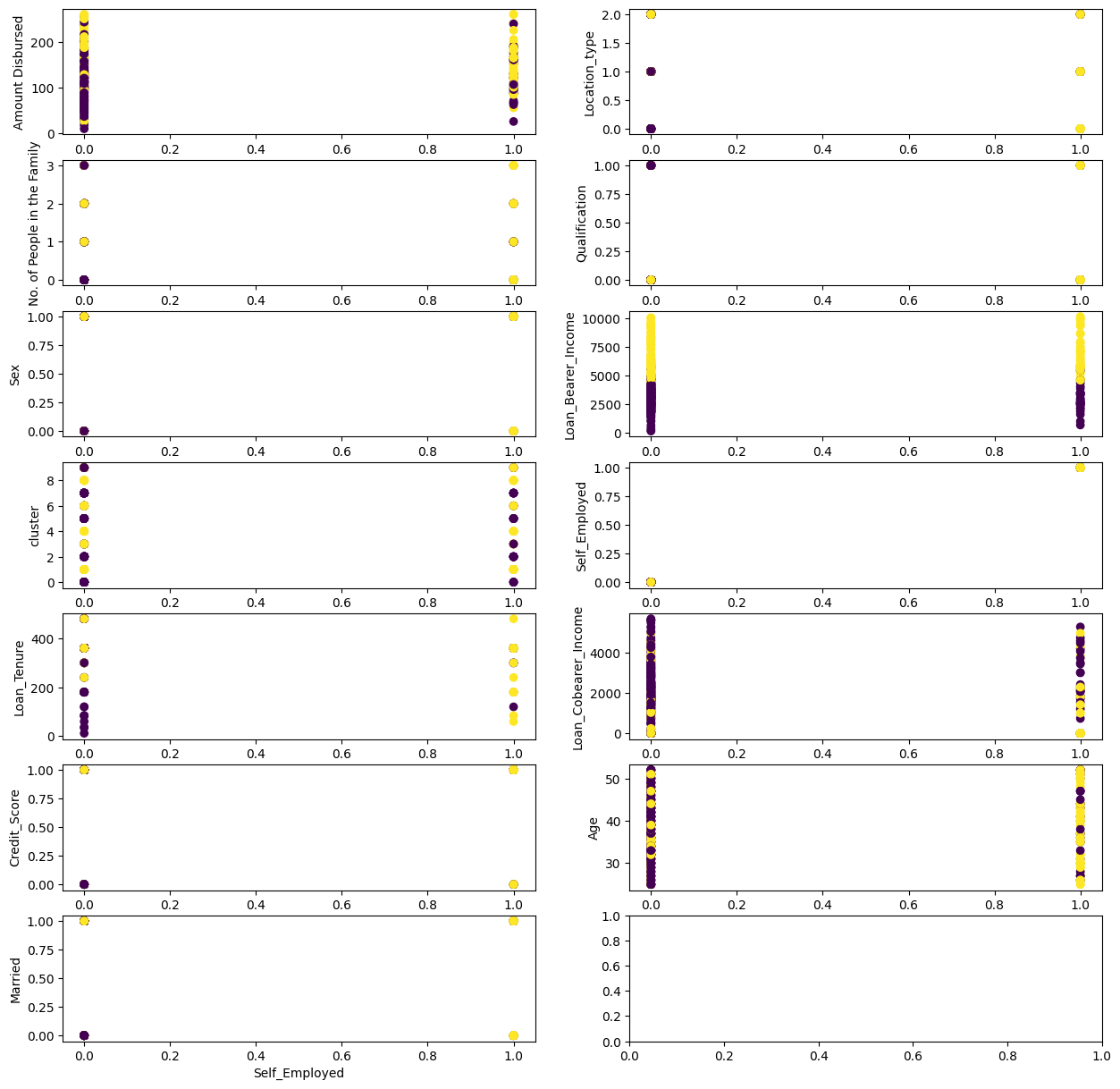
1. **Age**
2. **Married:**
3. **No of people in family:**

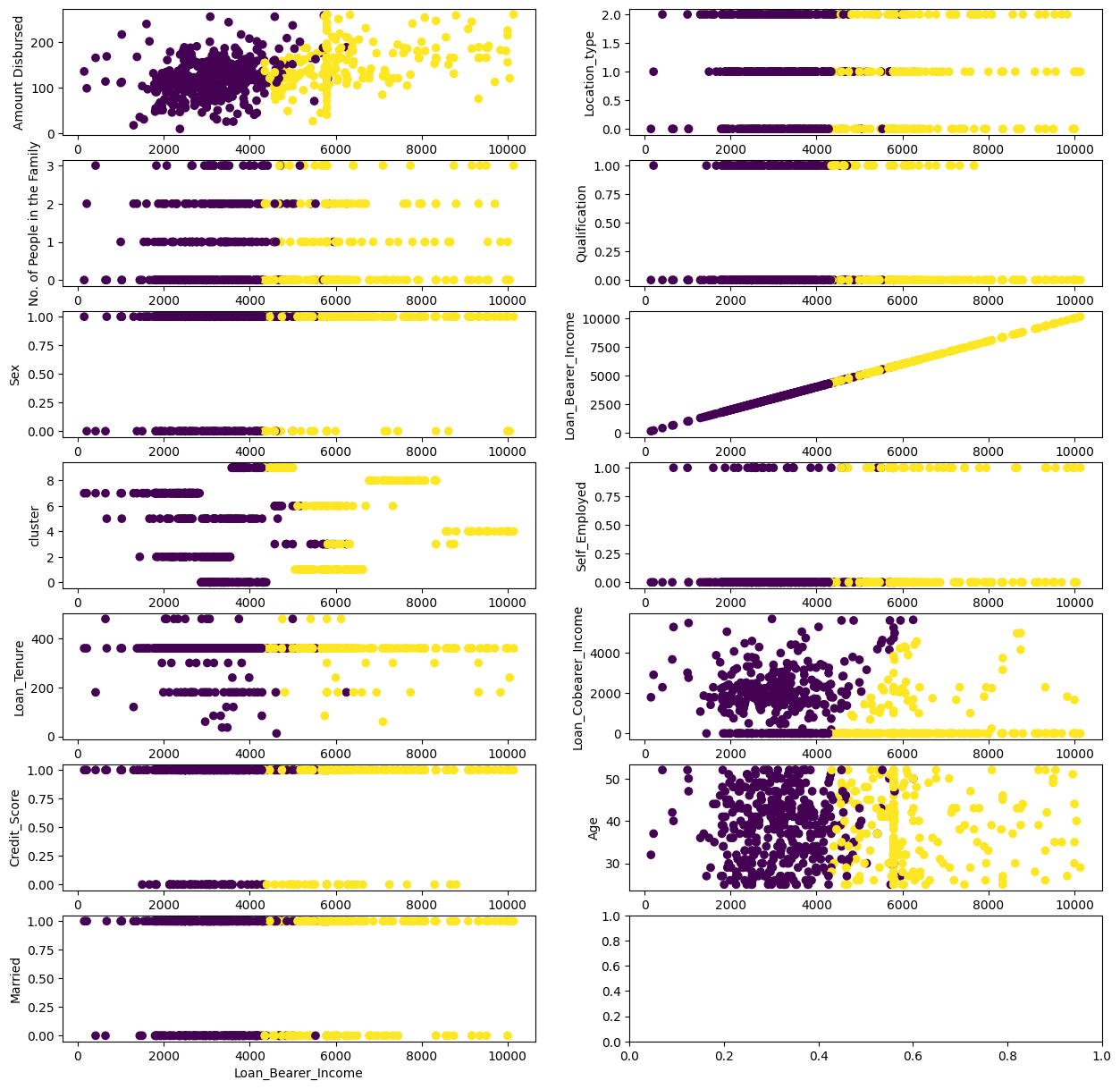
****

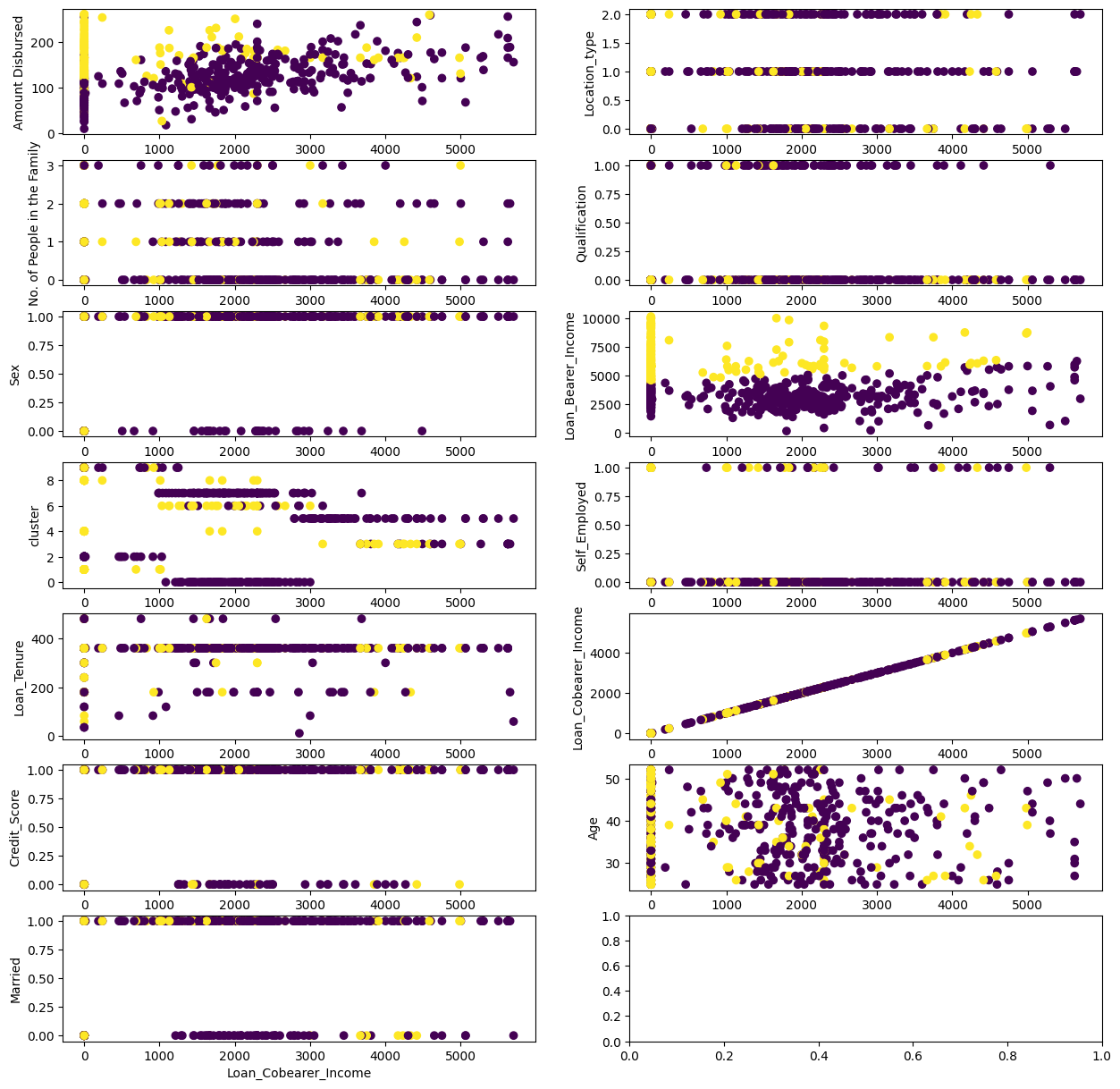
1. **Qualification:**

****

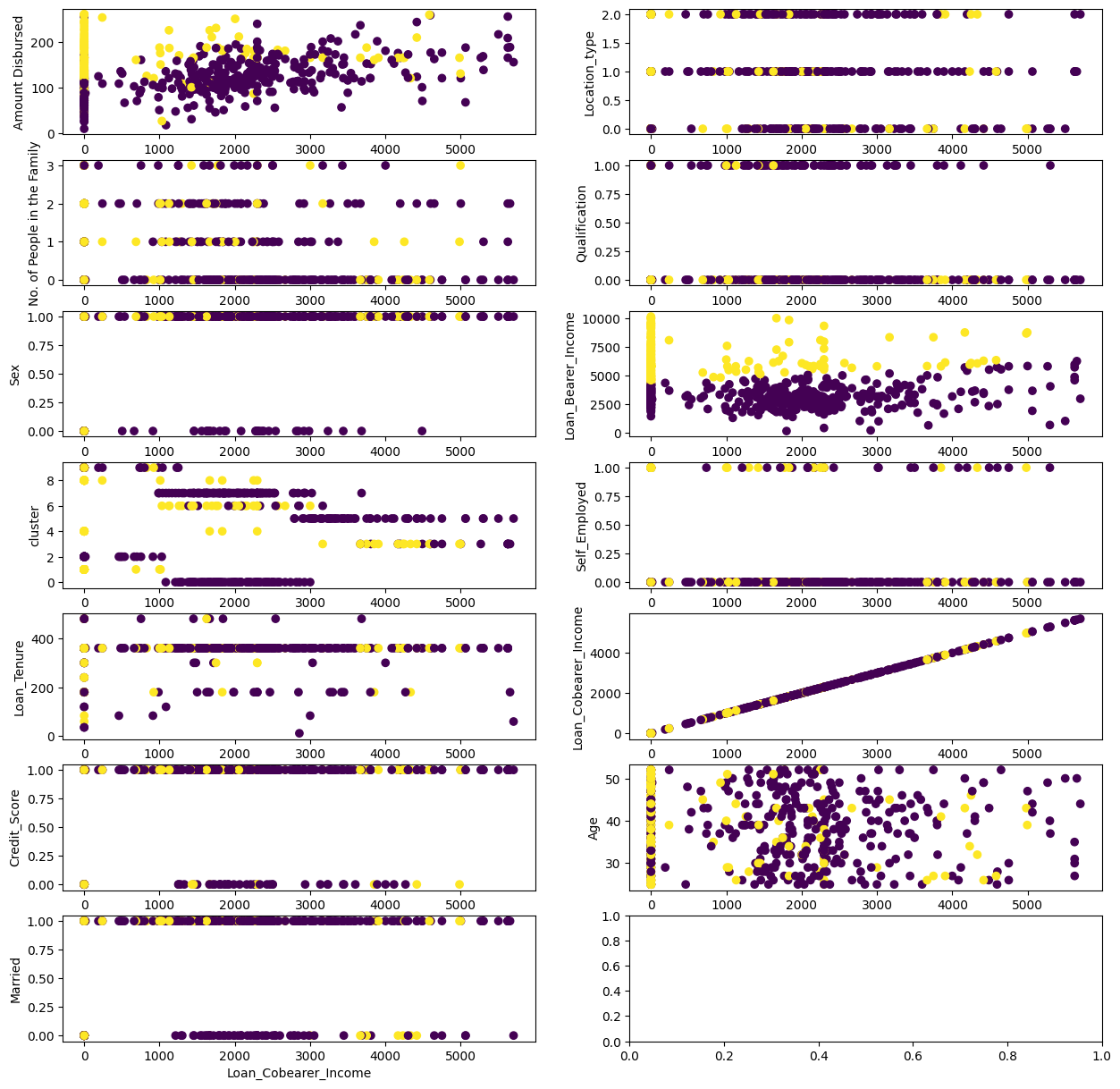
1. **Self employed:**

****

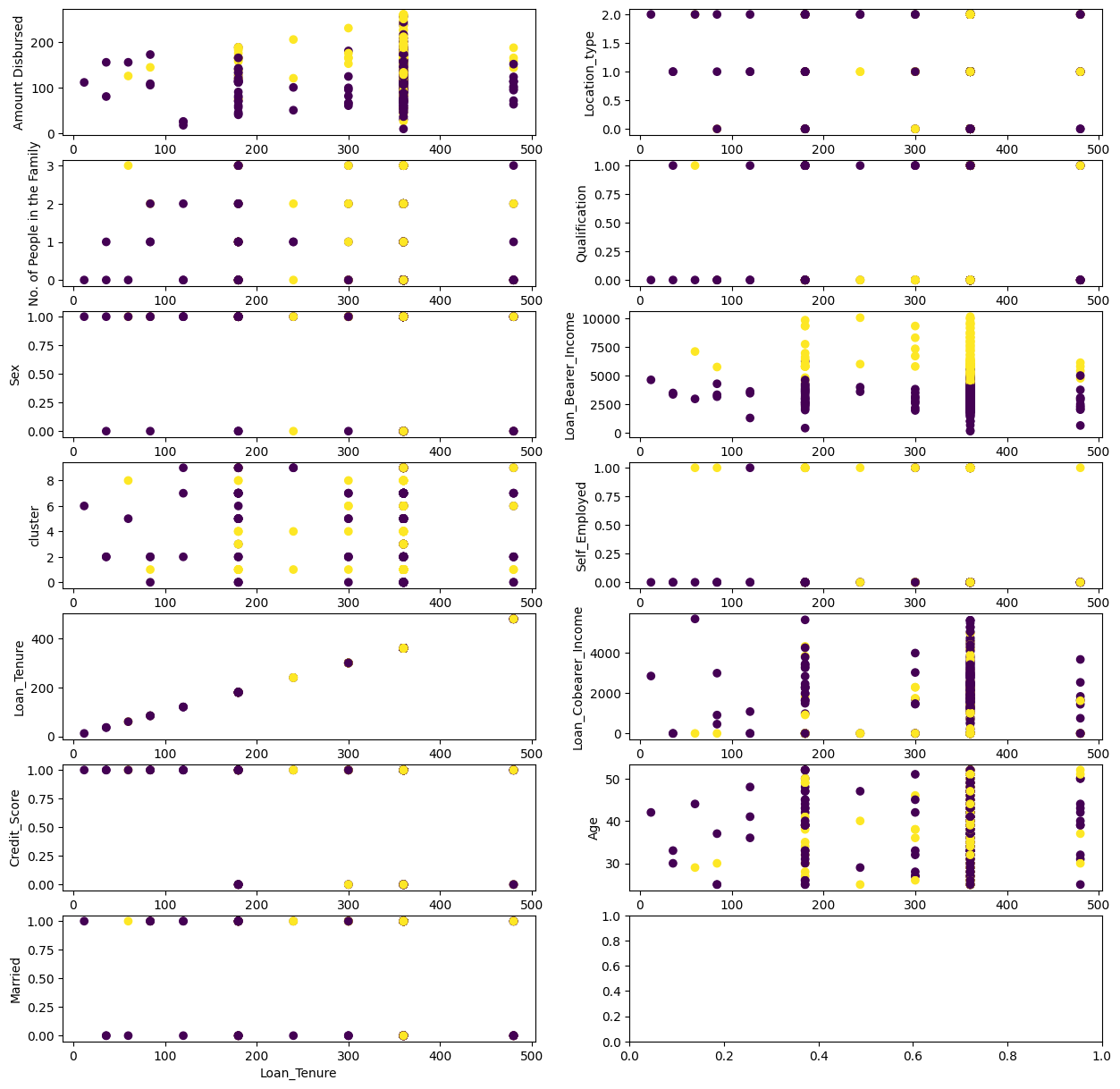
1. **Loan bearer income:**
2. **Loan Cobearer income:**

****

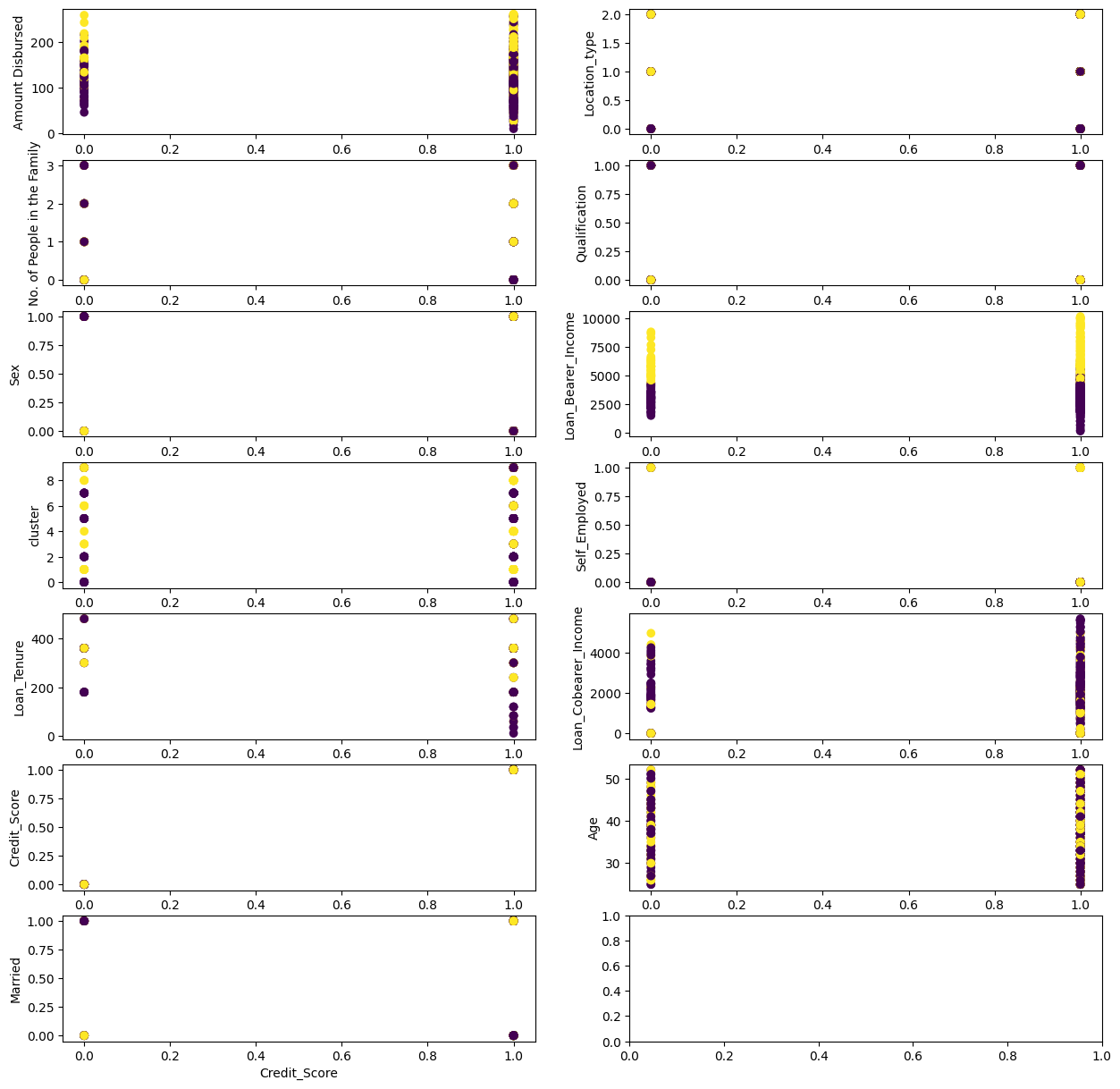
1. **Amount disbursed:**

****

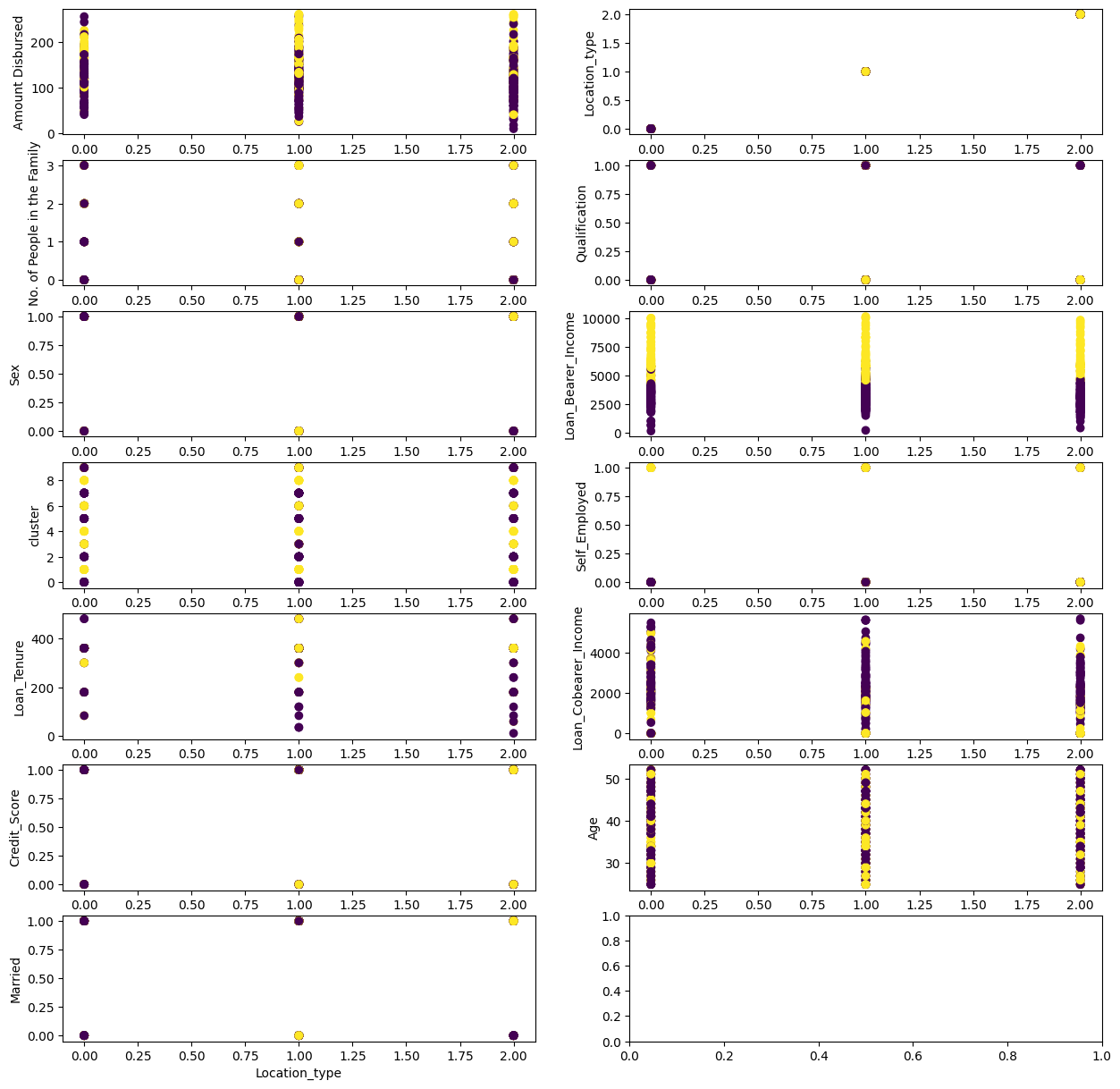
1. **Loan tenure:**

****

1. **Credit score:**

****

1. **Location type:**

****

**Analysis:**

* Loan bearer income is forming clear clusters using all the independent variables.
* In column cluster, 1 represents data points belonging to cluster 1, 0 represents data points belonging to cluster 0.
* As values of cluster are less in scale compared to other values in different columns, we are trying to come up with a value which will we in range with respect to other value.
* In this case as cement is forming clear cluster, we are using that column.
* For this purpose, we have grouped cluster column with cement column and found mean and median of their respective cluster.

Using Clusters to improve accuracy:

**Model building and evaluation:**

**Normal model:**

|  |  |  |
| --- | --- | --- |
|  | model\_name | accuracy\_score |
| 0 | LogisticRegression | 0.789189 |
| 4 | SVC | 0.789189 |
| 3 | RandomForestClassifier | 0.762162 |
| 2 | KneighborClassifier | 0.756757 |
| 6 | GradientBoostingClassifier | 0.756757 |
| 7 | XGBClassifier | 0.751351 |
| 5 | AdaBoostClassifier | 0.745946 |
| 1 | DecisionTreeClassifier | 0.686486 |

**Applying Cross validation:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | model\_names | cv\_score | cv\_std |
| 4 | SVC | 0.809572 | 0.033478 |
| 0 | Logisticregression | 0.807932 | 0.031692 |
| 6 | GradientBoostingClassifier | 0.781861 | 0.053825 |
| 2 | KNeighborsClassifier | 0.780328 | 0.044588 |
| 3 | RandomForestClassifier | 0.775278 | 0.050187 |
| 5 | AdaBoostClassifier | 0.760762 | 0.082982 |
| 7 | XGBClassifier | 0.753966 | 0.062846 |
| 1 | DecisionTreeClassifier | 0.682364 | 0.049036 |

**Cross Validation post hyperparameter:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | model\_names | cv\_score | cv\_std |
| 0 | LogisticRegression | 0.807932 | 0.031692 |
| 1 | RandomForestClassifier | 0.803014 | 0.029859 |
| 5 | AdaBoostClassifier | 0.790005 | 0.05186 |
| 6 | GradientBoostingClassifier | 0.788313 | 0.047107 |
| 4 | SVC | 0.781809 | 0.027028 |
| 3 | KNeighborsClassifier | 0.780328 | 0.044588 |
| 2 | DecisionTreeClassifier | 0.755685 | 0.041853 |
| 7 | XGBClassifier | 0.713406 | 0.071778 |

**Model building and evaluation: Cluster data**

**Normal model:**

|  |  |  |
| --- | --- | --- |
|  | model\_name | accuracy\_score |
| 6 | GradientBoostingClassifier | 0.794595 |
| 5 | AdaBoostClassifier | 0.783784 |
| 3 | RandomForestClassifier | 0.772973 |
| 7 | XGBClassifier | 0.772973 |
| 1 | DecisionTreeClassifier | 0.713514 |
| 0 | LogisticRegression | 0.67027 |
| 4 | SVC | 0.67027 |
| 2 | KneighborClassifier | 0.654054 |

**Applying Cross validation:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | model\_names | cv\_score | cv\_std |
| 6 | GradientBoostingClassifier | 0.791645 | 0.038461 |
| 3 | RandomForestClassifier | 0.785167 | 0.037495 |
| 5 | AdaBoostClassifier | 0.785061 | 0.029624 |
| 7 | XGBClassifier | 0.755843 | 0.048715 |
| 1 | DecisionTreeClassifier | 0.705315 | 0.052556 |
| 0 | Logisticregression | 0.687308 | 0.020112 |
| 4 | SVC | 0.687308 | 0.005314 |
| 2 | KNeighborsClassifier | 0.631994 | 0.024236 |

**Cross Validation post hyperparameter:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | model\_names | cv\_score | cv\_std |
| 5 | AdaBoostClassifier | 0.799788 | 0.032693 |
| 6 | GradientBoostingClassifier | 0.796483 | 0.036992 |
| 2 | DecisionTreeClassifier | 0.768773 | 0.053705 |
| 1 | RandomForestClassifier | 0.763934 | 0.049545 |
| 7 | XGBClassifier | 0.711846 | 0.046364 |
| 0 | LogisticRegression | 0.687308 | 0.020112 |
| 4 | SVC | 0.687308 | 0.005314 |
| 3 | KNeighborsClassifier | 0.631994 | 0.024236 |

**Model building and evaluation**: After feature selection and following features were eliminated 'Sex', 'Qualification', 'Self\_Employed', 'cluster', 'mean', 'median'

**Normal model:**

|  |  |  |
| --- | --- | --- |
|  | model\_name | accuracy\_score |
| 6 | GradientBoostingClassifier | 0.794595 |
| 0 | LogisticRegression | 0.783784 |
| 3 | RandomForestClassifier | 0.783784 |
| 5 | AdaBoostClassifier | 0.772973 |
| 1 | DecisionTreeClassifier | 0.756757 |
| 7 | XGBClassifier | 0.756757 |
| 4 | SVC | 0.67027 |
| 2 | KneighborClassifier | 0.643243 |

**Applying Cross validation:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | model\_names | cv\_score | cv\_std |
| 0 | Logisticregression | 0.801375 | 0.028415 |
| 6 | GradientBoostingClassifier | 0.788366 | 0.043928 |
| 5 | AdaBoostClassifier | 0.784955 | 0.027542 |
| 3 | RandomForestClassifier | 0.780328 | 0.043366 |
| 7 | XGBClassifier | 0.773718 | 0.045958 |
| 1 | DecisionTreeClassifier | 0.710259 | 0.052087 |
| 4 | SVC | 0.687308 | 0.005314 |
| 2 | KNeighborsClassifier | 0.630381 | 0.024941 |

**Cross Validation post hyperparameter:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | model\_names | cv\_score | cv\_std |
| 1 | RandomForestClassifier | 0.809572 | 0.033478 |
| 0 | LogisticRegression | 0.801375 | 0.028415 |
| 2 | DecisionTreeClassifier | 0.799788 | 0.035171 |
| 5 | AdaBoostClassifier | 0.799788 | 0.034298 |
| 6 | GradientBoostingClassifier | 0.793231 | 0.036828 |
| 7 | XGBClassifier | 0.713432 | 0.04988 |
| 4 | SVC | 0.687308 | 0.005314 |
| 3 | KNeighborsClassifier | 0.630381 | 0.024941 |

**Model building and evaluation**: After feature selection and following features were eliminated 'Sex', 'Married', 'No. of People in the Family', 'Loan\_Bearer\_Income', 'Amount Disbursed', 'cluster', 'mean', 'median'

**Normal model:**

|  |  |  |
| --- | --- | --- |
|  | model\_name | accuracy\_score |
| 5 | AdaBoostClassifier | 0.794595 |
| 0 | LogisticRegression | 0.789189 |
| 6 | GradientBoostingClassifier | 0.789189 |
| 7 | XGBClassifier | 0.756757 |
| 3 | RandomForestClassifier | 0.751351 |
| 1 | DecisionTreeClassifier | 0.702703 |
| 2 | KneighborClassifier | 0.681081 |
| 4 | SVC | 0.67027 |

**Applying Cross validation:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | model\_names | cv\_score | cv\_std |
| 0 | Logisticregression | 0.809572 | 0.033478 |
| 6 | GradientBoostingClassifier | 0.796616 | 0.042857 |
| 5 | AdaBoostClassifier | 0.791618 | 0.041749 |
| 3 | RandomForestClassifier | 0.757403 | 0.046652 |
| 7 | XGBClassifier | 0.751137 | 0.062612 |
| 4 | SVC | 0.687308 | 0.005314 |
| 1 | DecisionTreeClassifier | 0.676018 | 0.036292 |
| 2 | KNeighborsClassifier | 0.659651 | 0.049971 |

**Cross Validation post hyperparameter:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | model\_names | cv\_score | cv\_std |
| 6 | GradientBoostingClassifier | 0.809598 | 0.036321 |
| 0 | LogisticRegression | 0.809572 | 0.033478 |
| 1 | RandomForestClassifier | 0.809572 | 0.033478 |
| 2 | DecisionTreeClassifier | 0.807959 | 0.043509 |
| 5 | AdaBoostClassifier | 0.803067 | 0.04083 |
| 4 | SVC | 0.687308 | 0.005314 |
| 3 | KNeighborsClassifier | 0.659651 | 0.049971 |
| 7 | XGBClassifier | 0.639503 | 0.168592 |