**REPORT**

The dataset was loaded and assessed for issues. The issues found are stated as following:

* Erroneous datatype of id.
* Erroneous datatype of SeriousDlqin2yrs.
* Presence of null values in no. dependents and monthly columns.

**Data cleaning:**

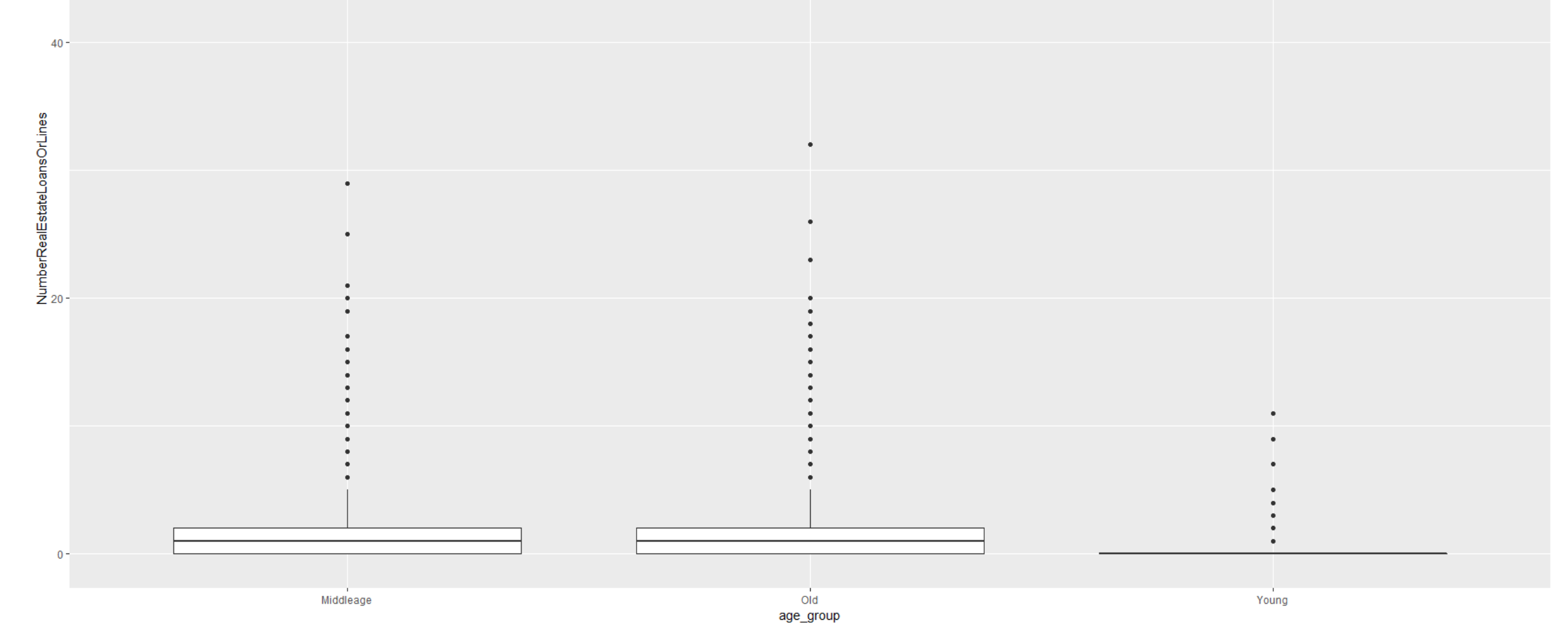
* The id datatype was converted to character.
* SeriousDlqin2years was converted to factors.
* The null value in no. dependents were replaced with 0.
* The null values for monthly income were replaced with mean monthly income value.
* A separate column called age\_group is created with following conditions:

1. age\_group=’young’ if age is between 18 and 30
2. age\_group=’middle age’ if age is between 30 and 55
3. age\_group=’old’ if age is greater than 55

**Exploratory Visualization:**

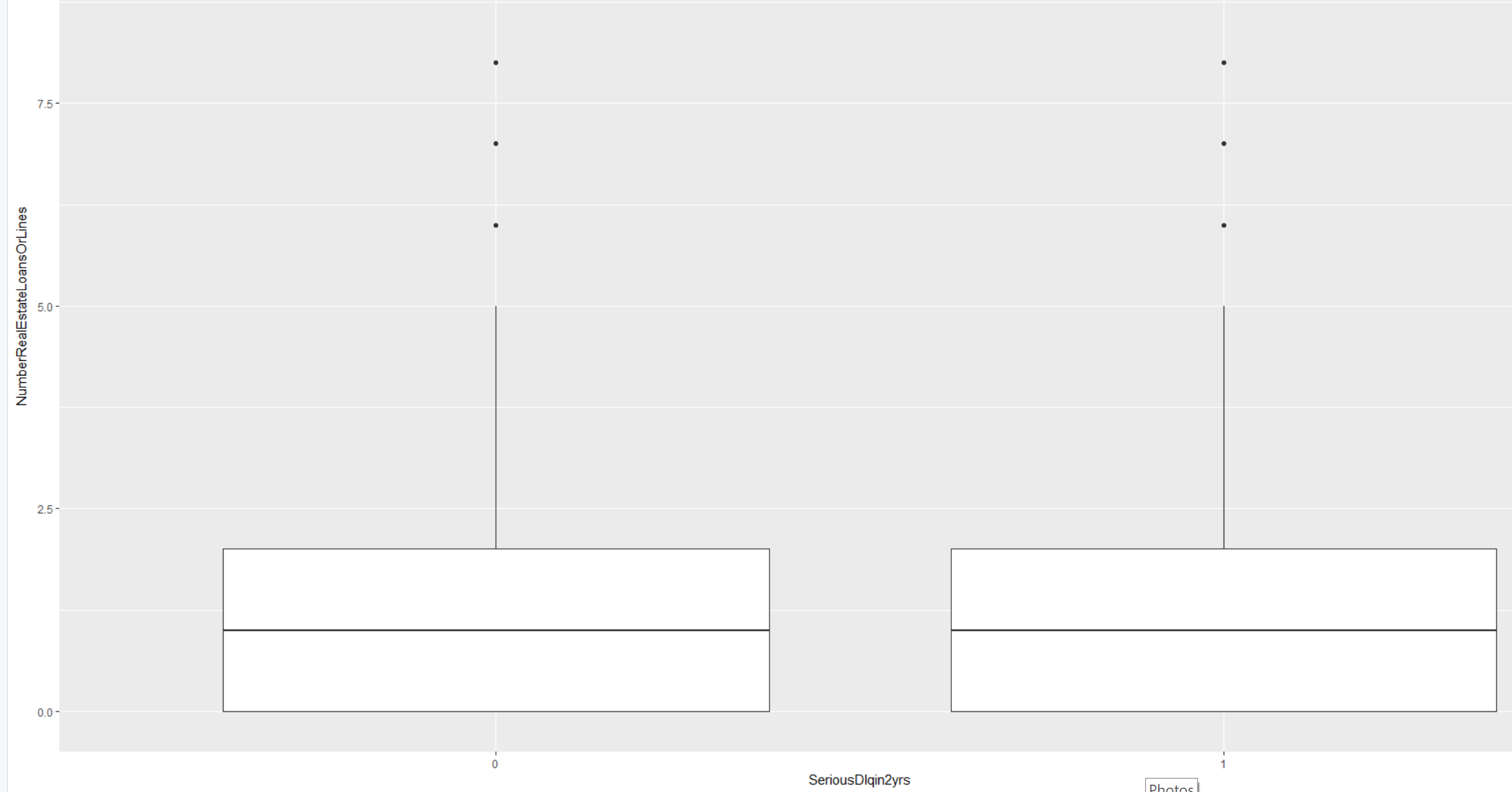
The following plots were plotted:

1. Boxplots
2. x=age\_group and y=NumberRealEstateLoansOrLines



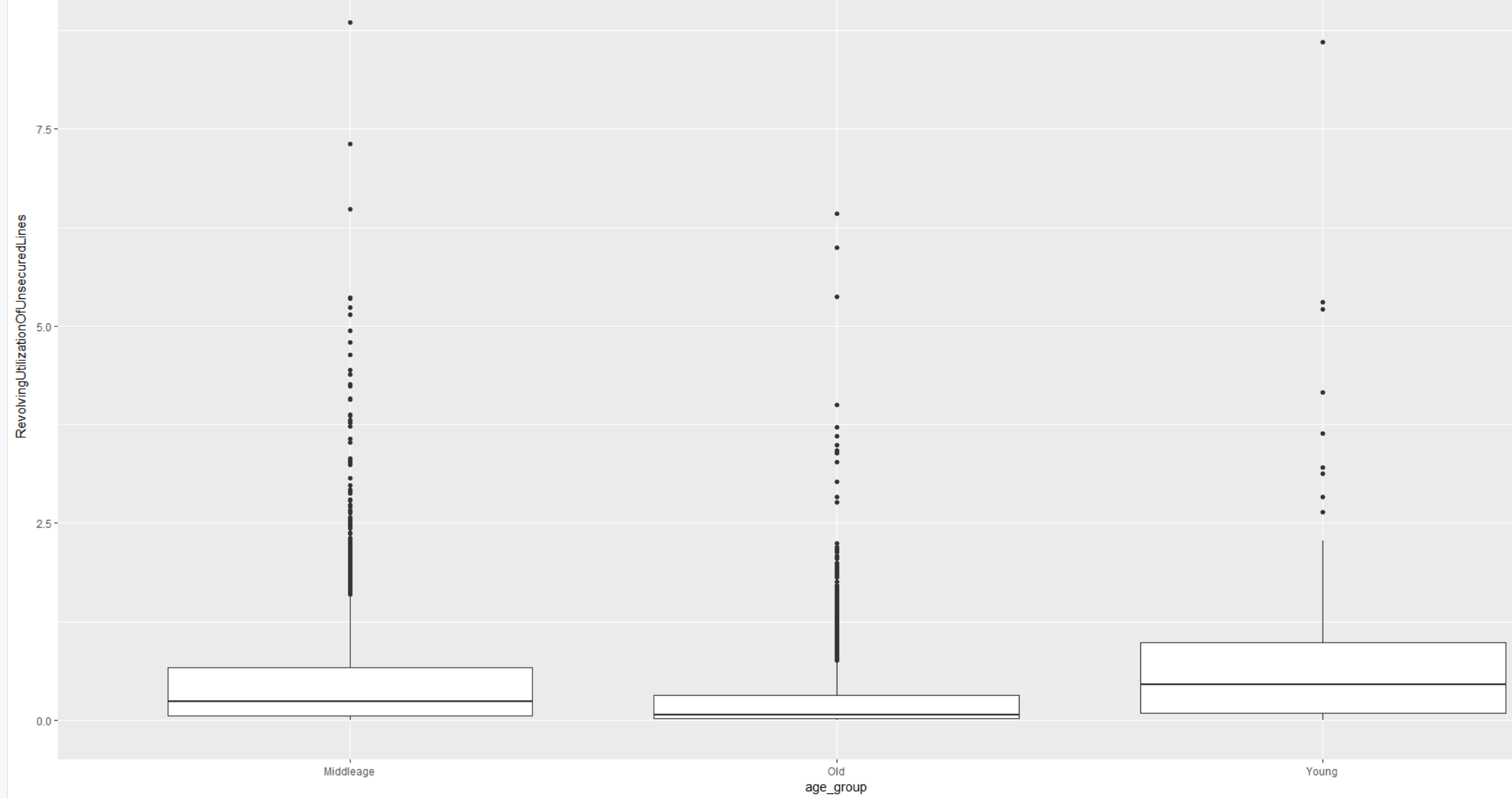
Old and middle age group people have almost the same range of real estate loans and lines and young people the least.

1. (x=SeriousDlqin2yrs, y=NumberRealEstateLoansOrLines)



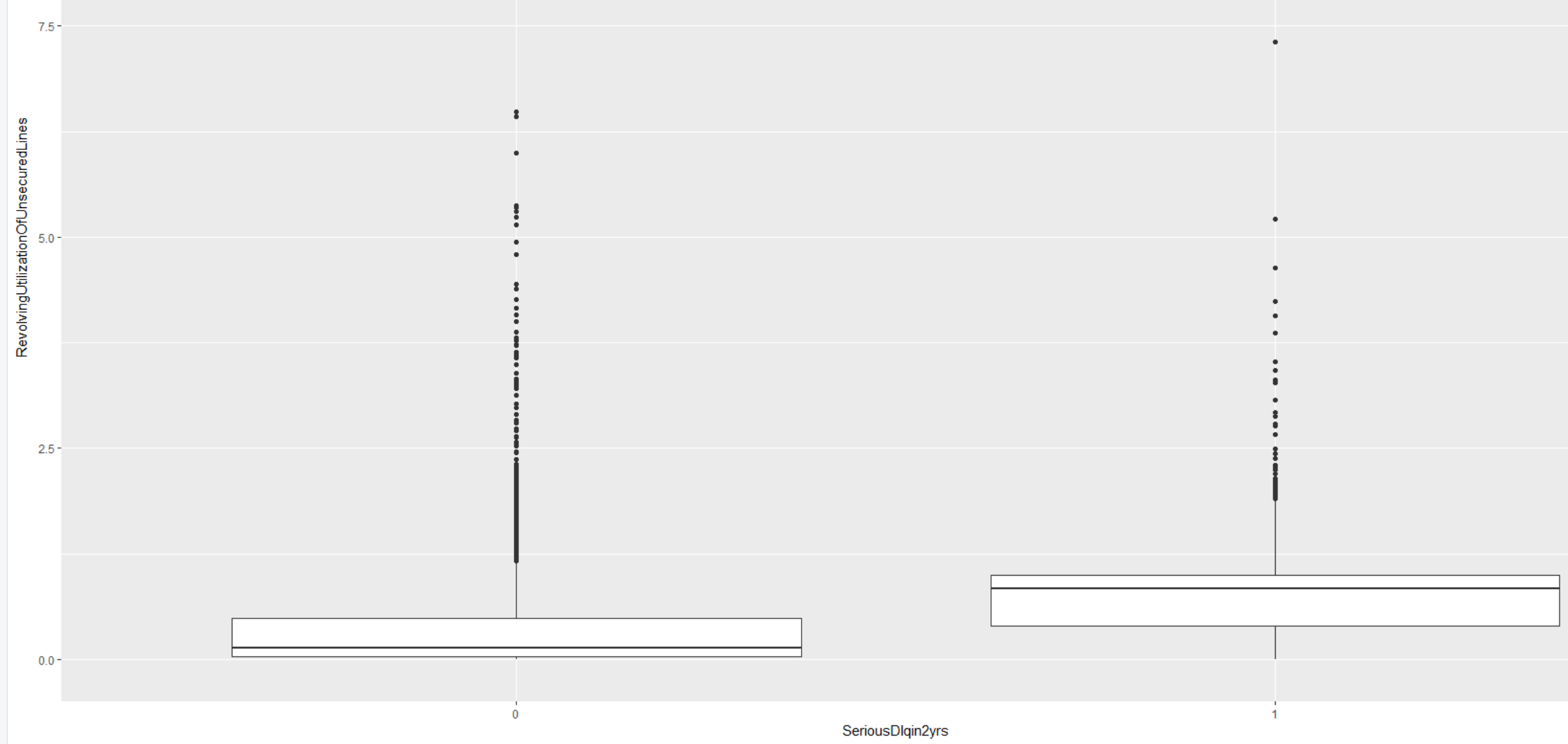
The number of real estate loans and lines are of the same range for both defaulters and non-defaulters.

1. (x=age\_group,y=RevolvingUtilizationOfUnsecuredLines)



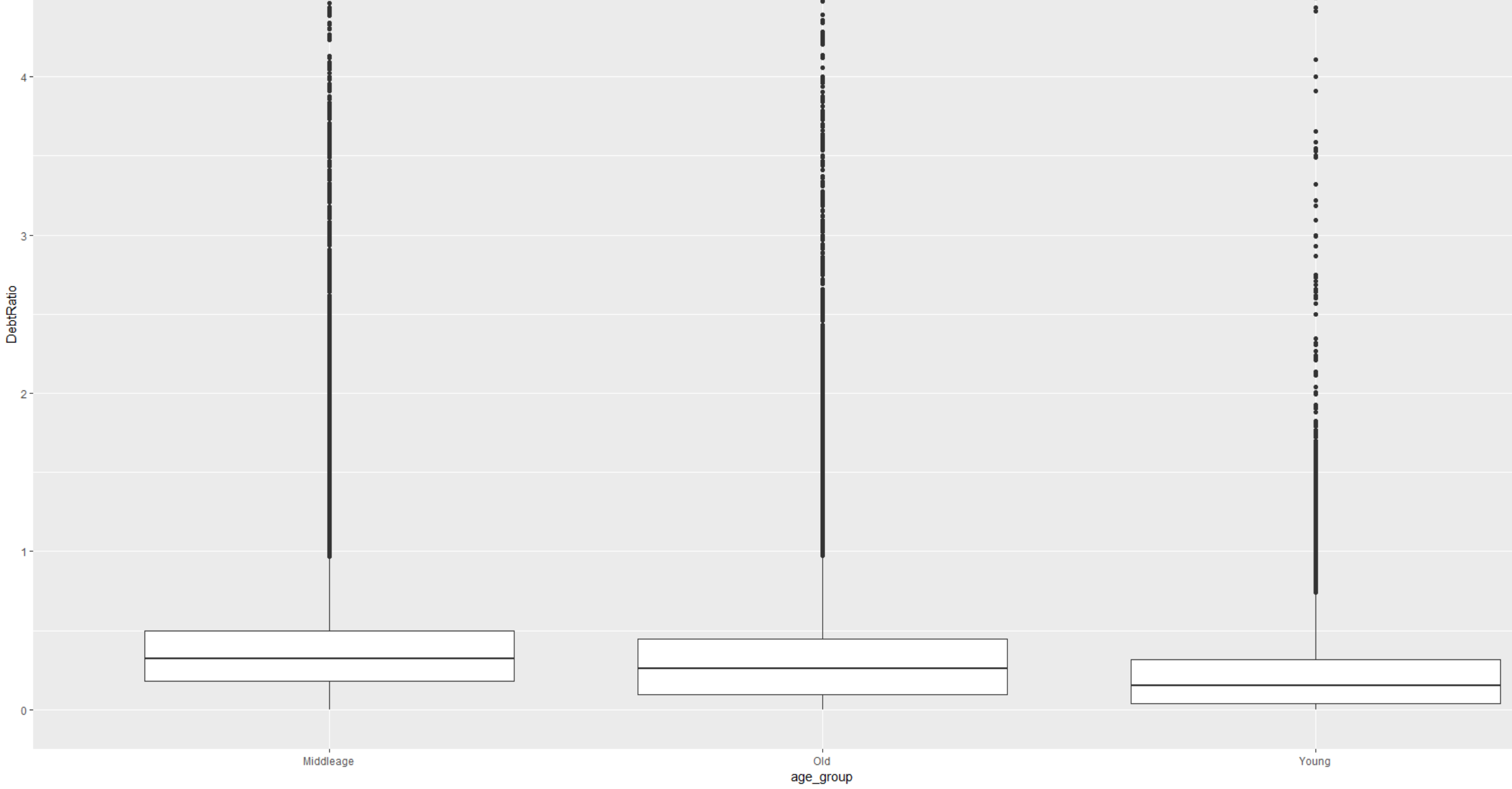
The median and the range of RevolvingUtilizationOfUnsecuredLine is higher for young people.

1. (x=SeriousDlqin2yrs, y=RevolvingUtilizationOfUnsecuredLines)



The median and the range of RevolvingUtilizationOfUnsecuredLine is higher for defaulters.

1. (x=age\_group,y=DebtRatio)



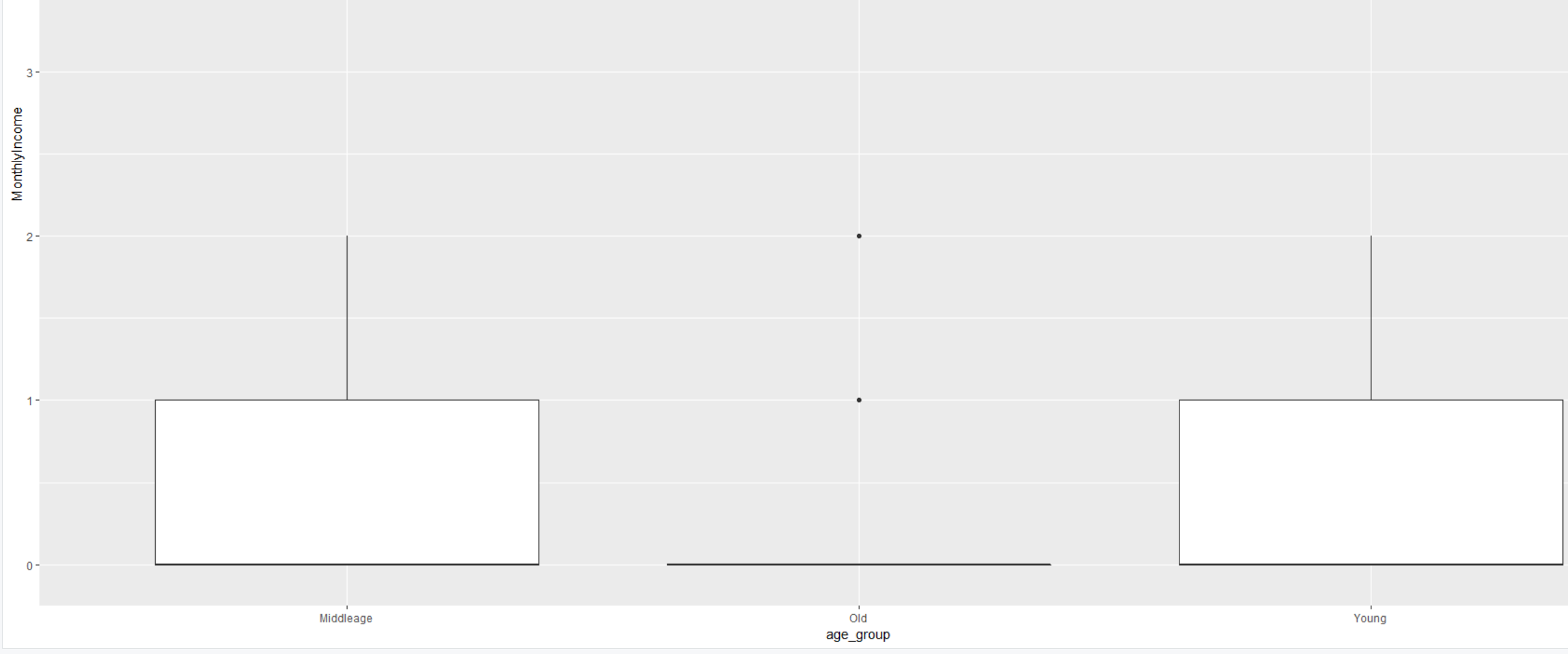
The debt ratio median is the highest for middle age and lowest for youngsters.

1. (x=SeriousDlqin2yrs, y=DebtRatio)



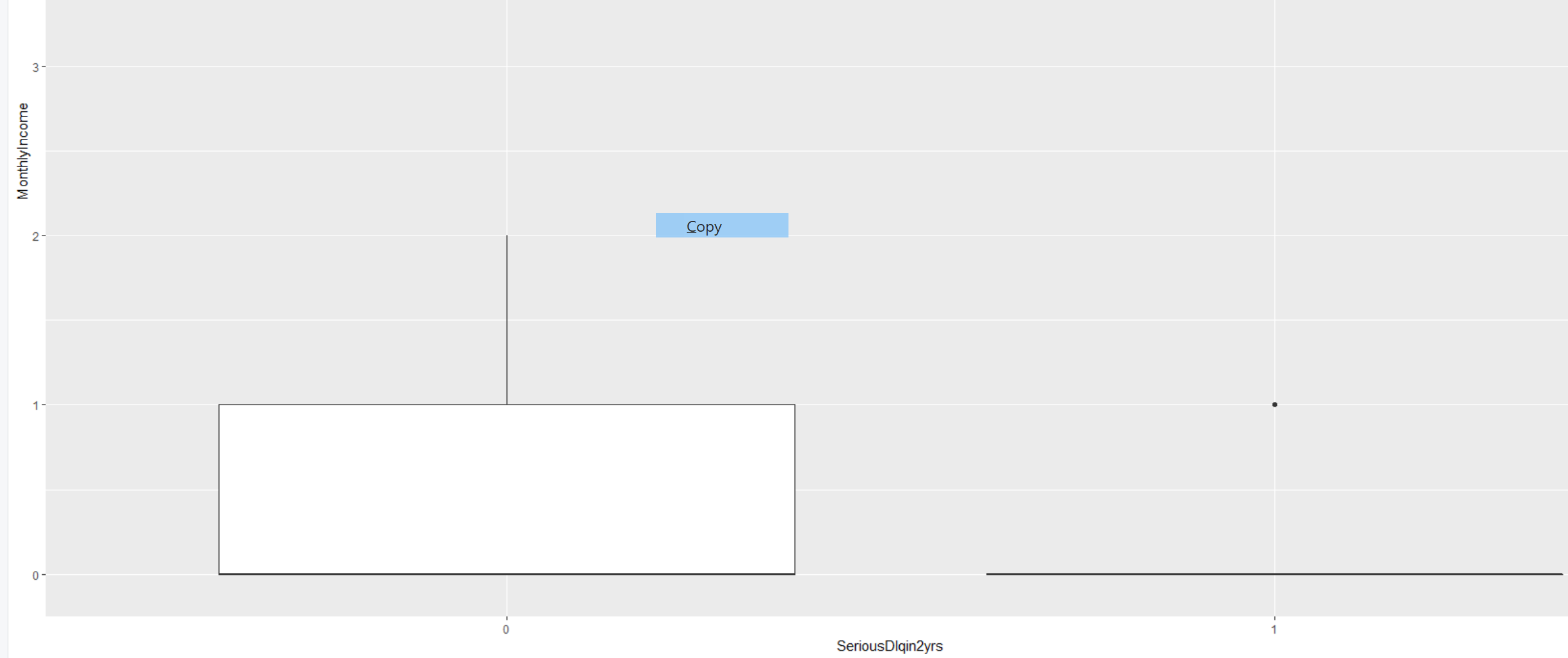
The median debt ratio for defaulters is higher than non-defaulters.

1. (x=age\_group, y=MonthlyIncome)



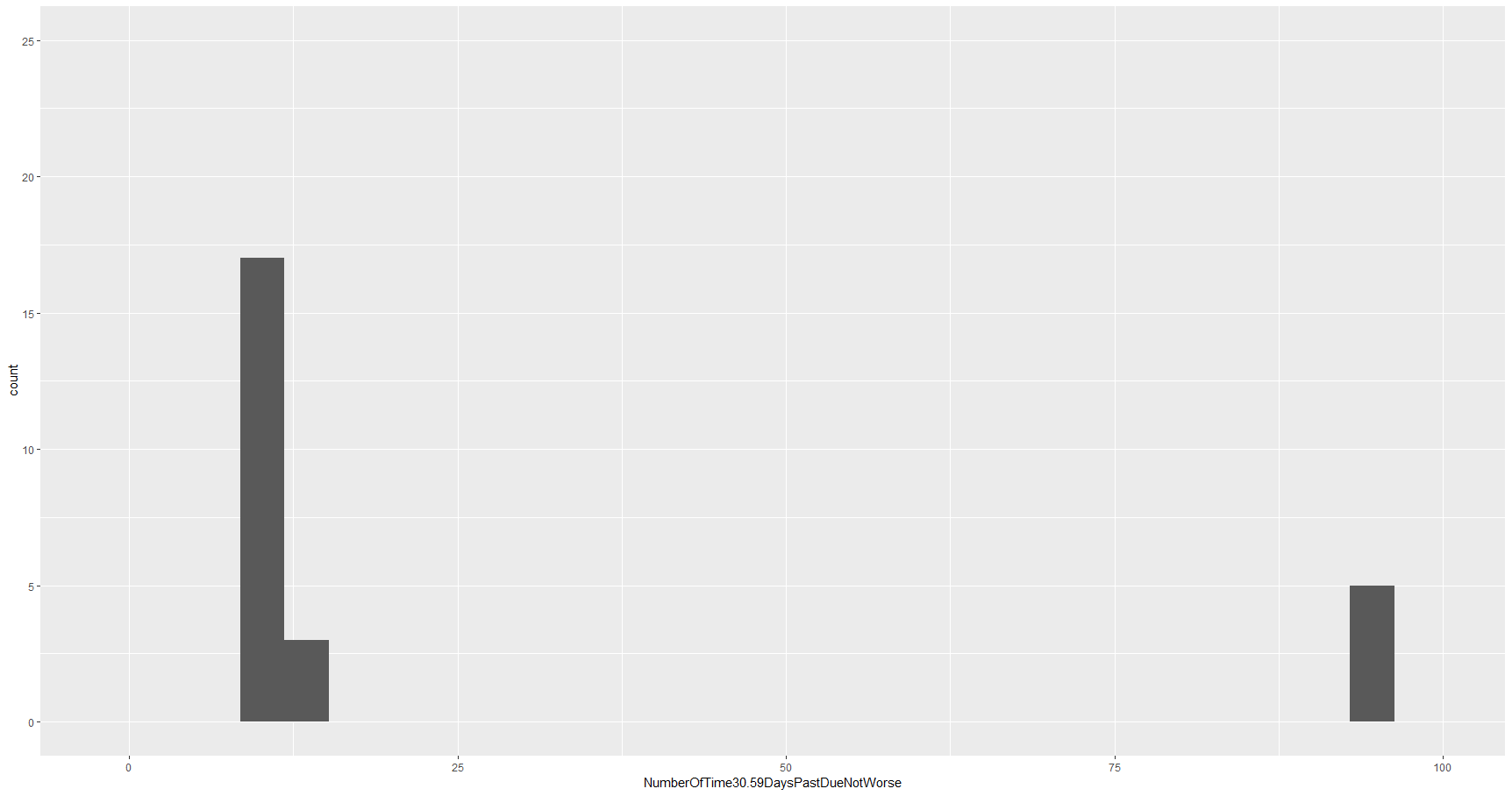
The monthly income range of both middleage and youngsters are almost the same and least for old people.

1. (x=SeriousDlqin2yrs,y=MonthlyIncome)



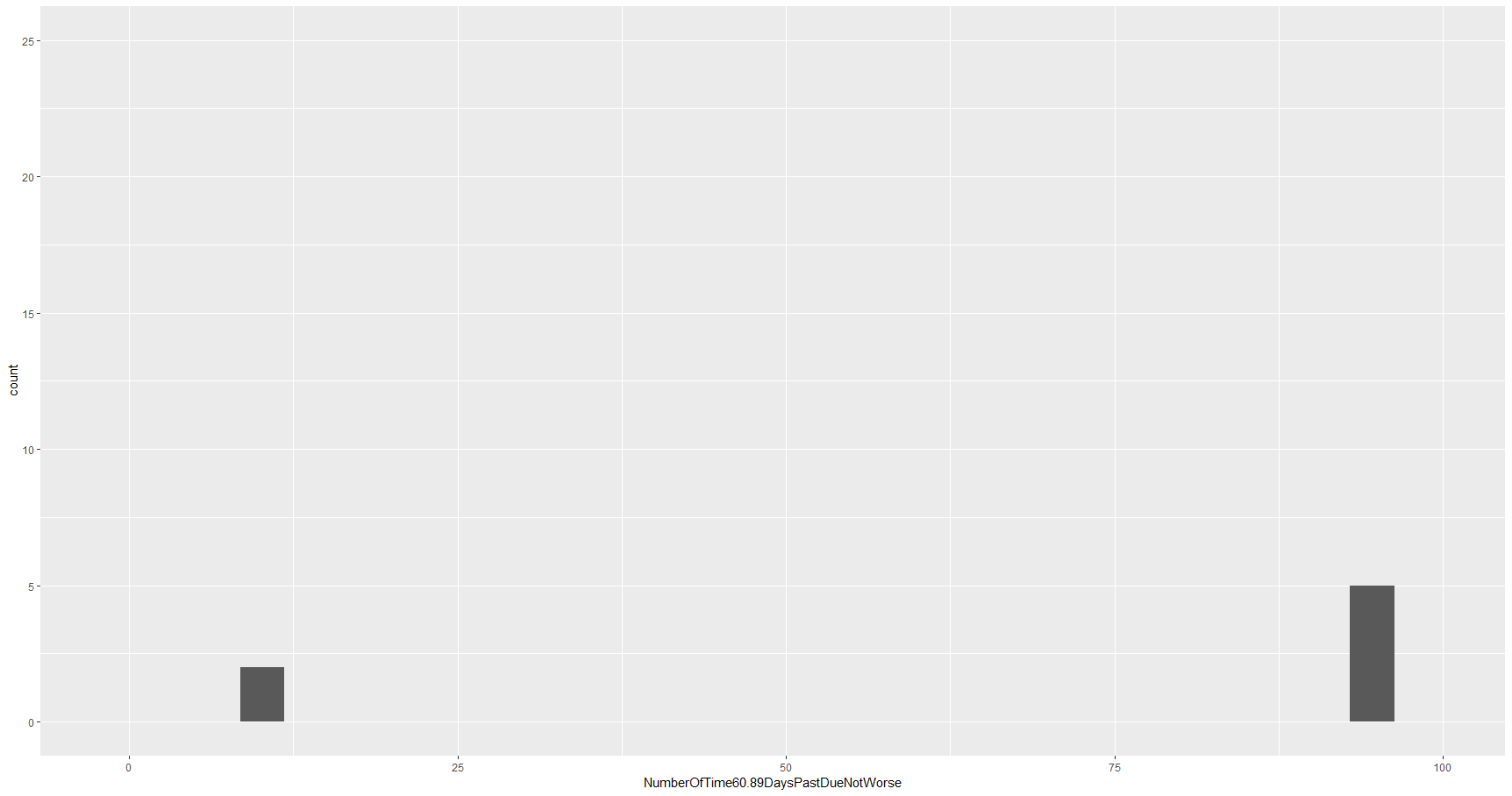
The monthly income of non-defaulters have a larger range.

1. Histogram
2. NumberOfTime30.59DaysPastDueNotWorse



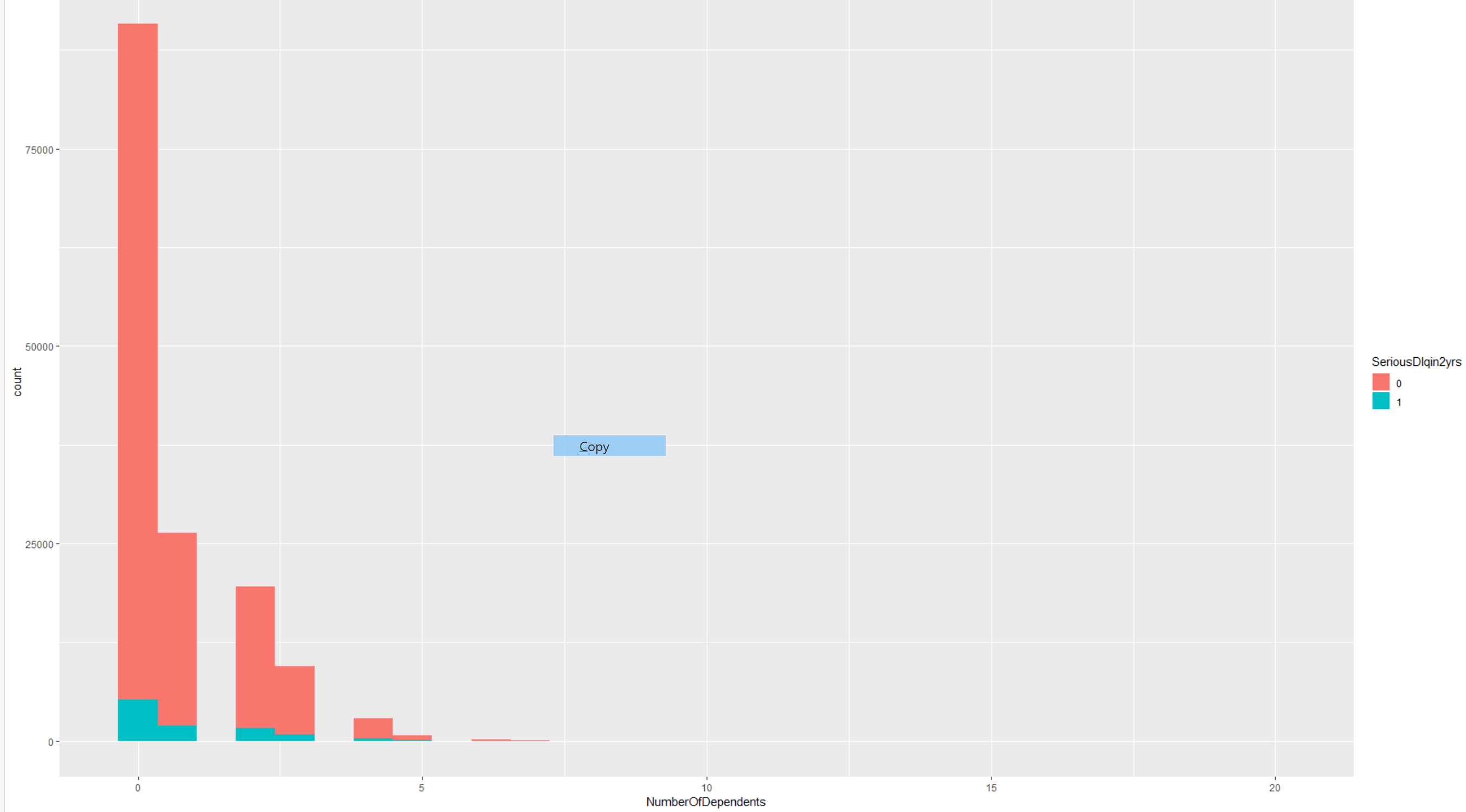
The frequency of number of times the payment was exceeded 30-59 times significantly at 96-98 and 7-13.

1. NumberOfTime60.89DaysPastDueNotWorse



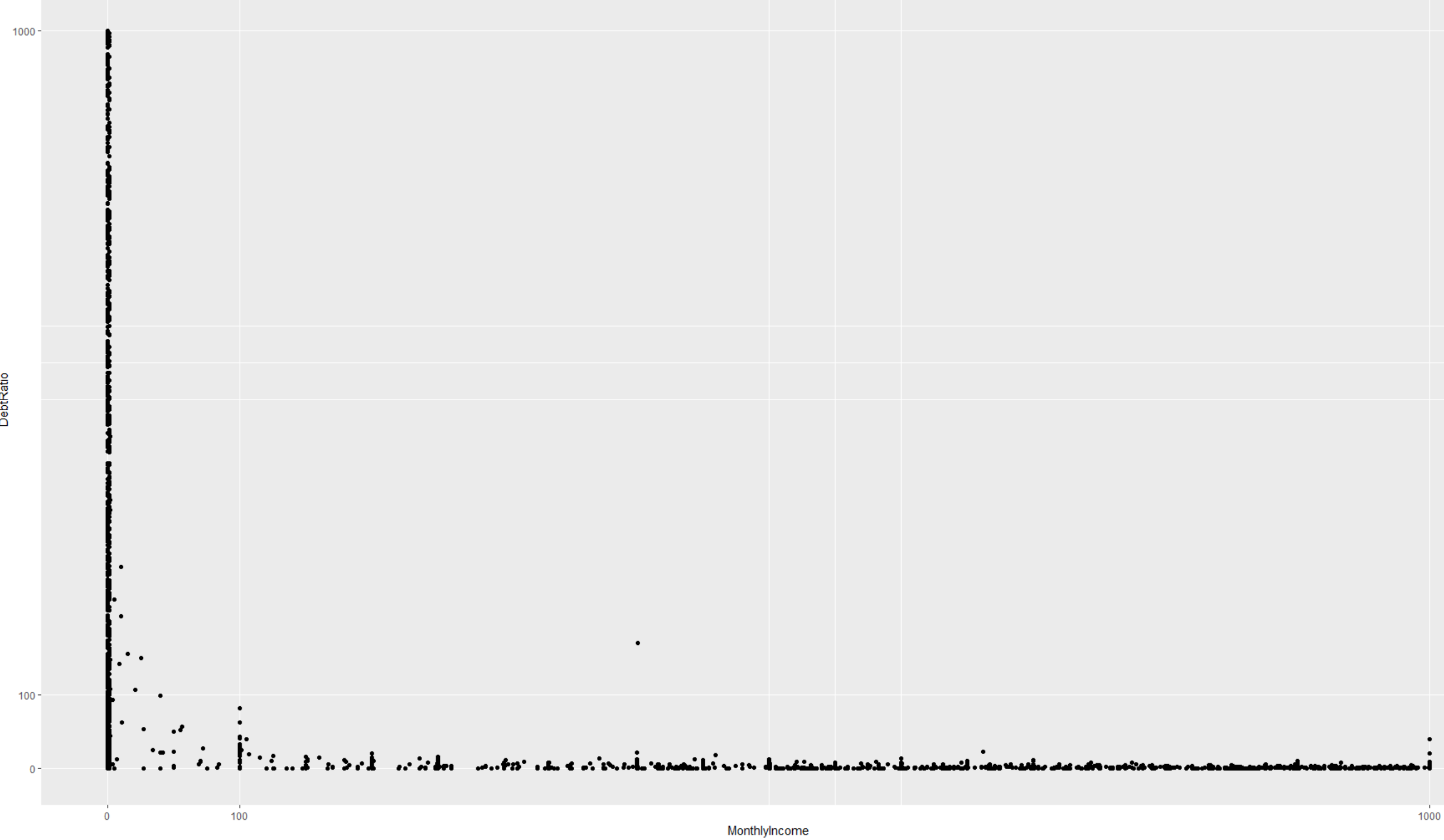
The frequency of number of times the payment was exceeded 60-89 times significantly at 6-11 and 96-98.

1. NumberOfDependents

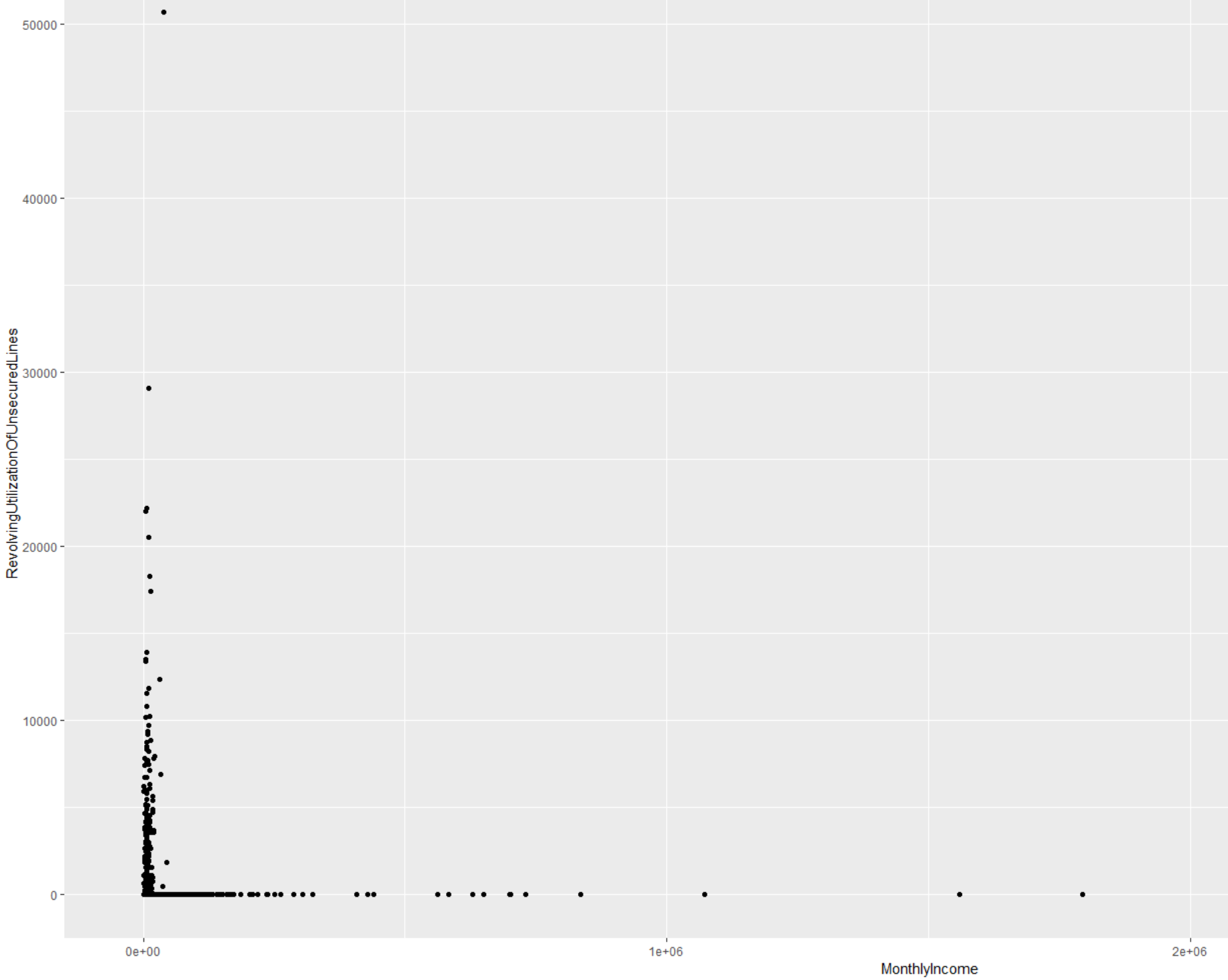


The number of dependents are much higher for non-defaulters than defaulters.

1. Scatterplot
2. (x=MonthlyIncome,y=DebtRatio)

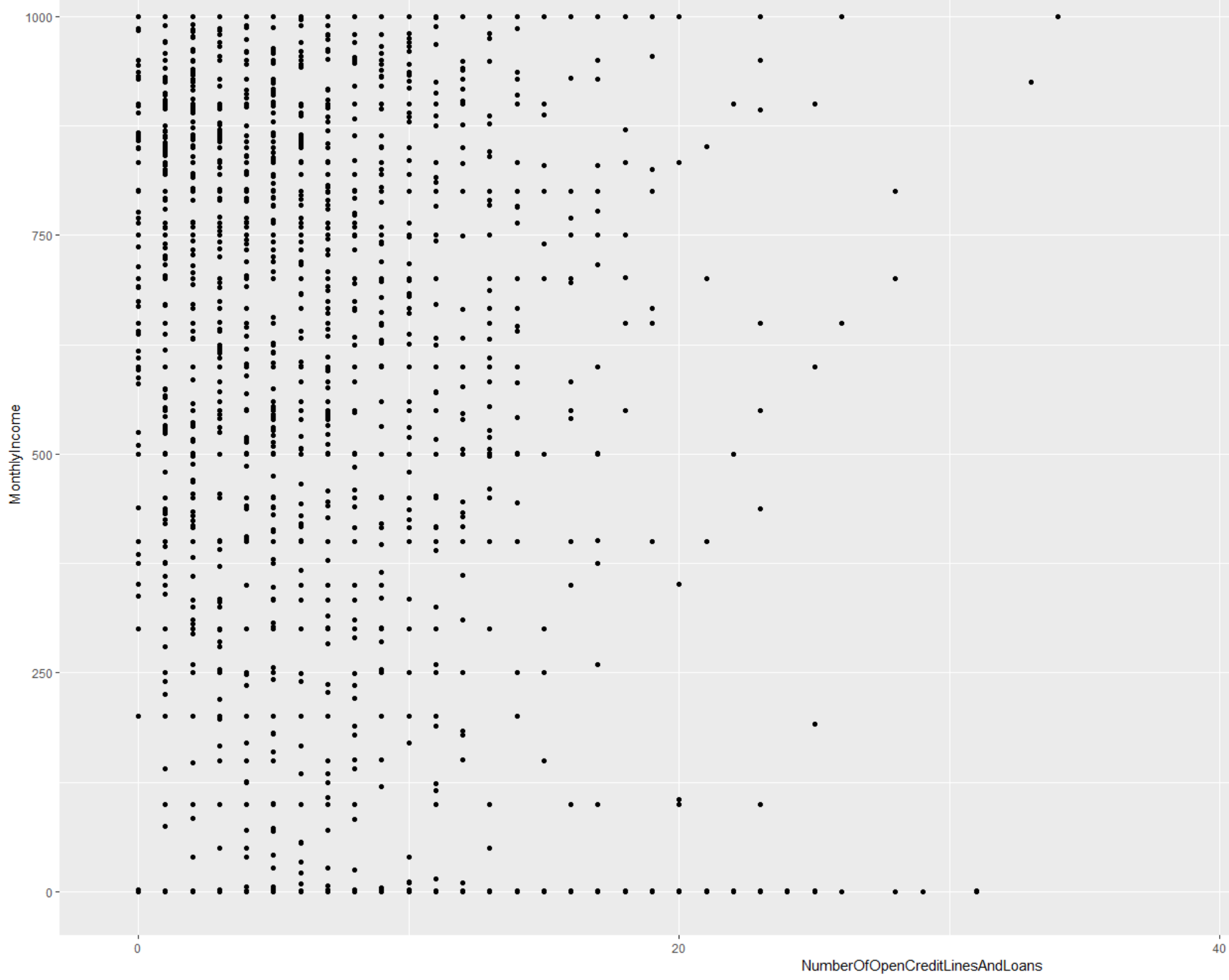
There is a very wide range of debt ratio and monthly income for low income and, as income decreases debt ratio also decreases.

1. (x=MonthlyIncome,y=RevolvingUtilizationOfUnsecuredLines)



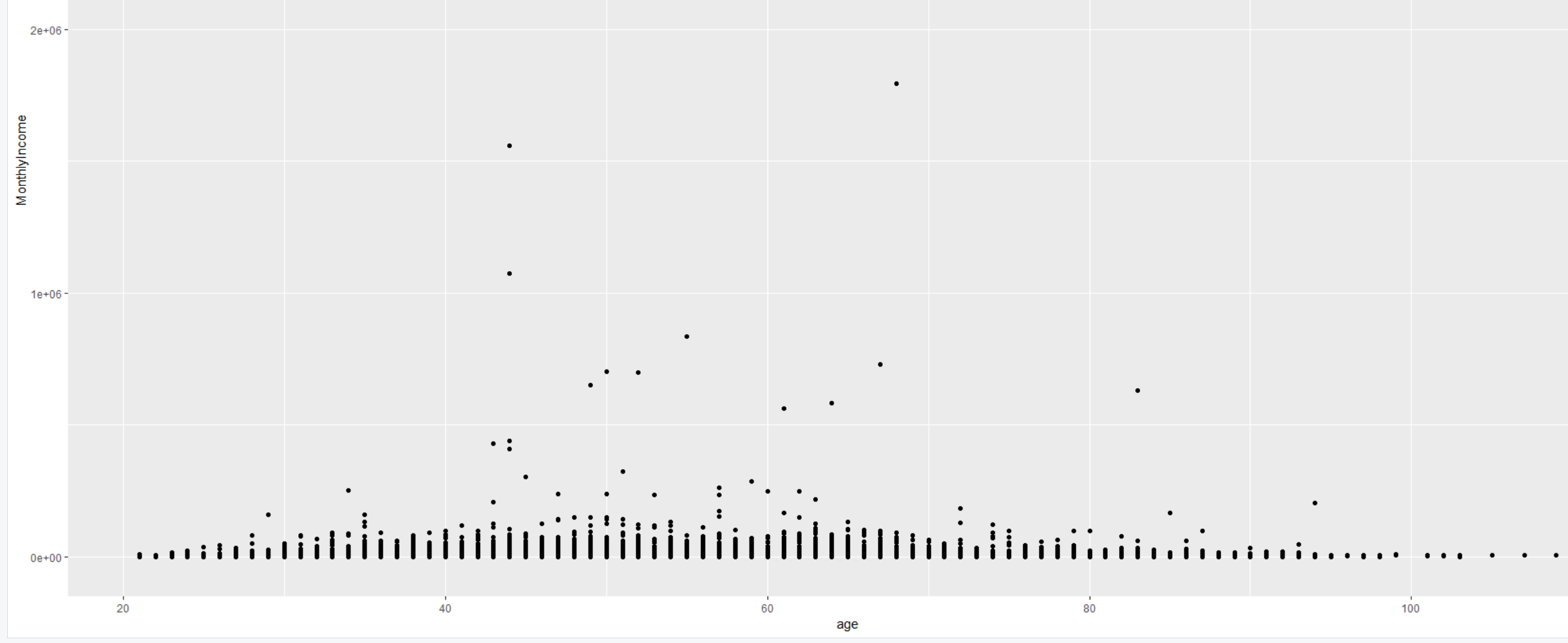
The revolving utilization of unsecured lines are more concentrated at low income.

1. (x= NumberOfOpenCreditLinesAndLoans,y=MonthlyIncome)



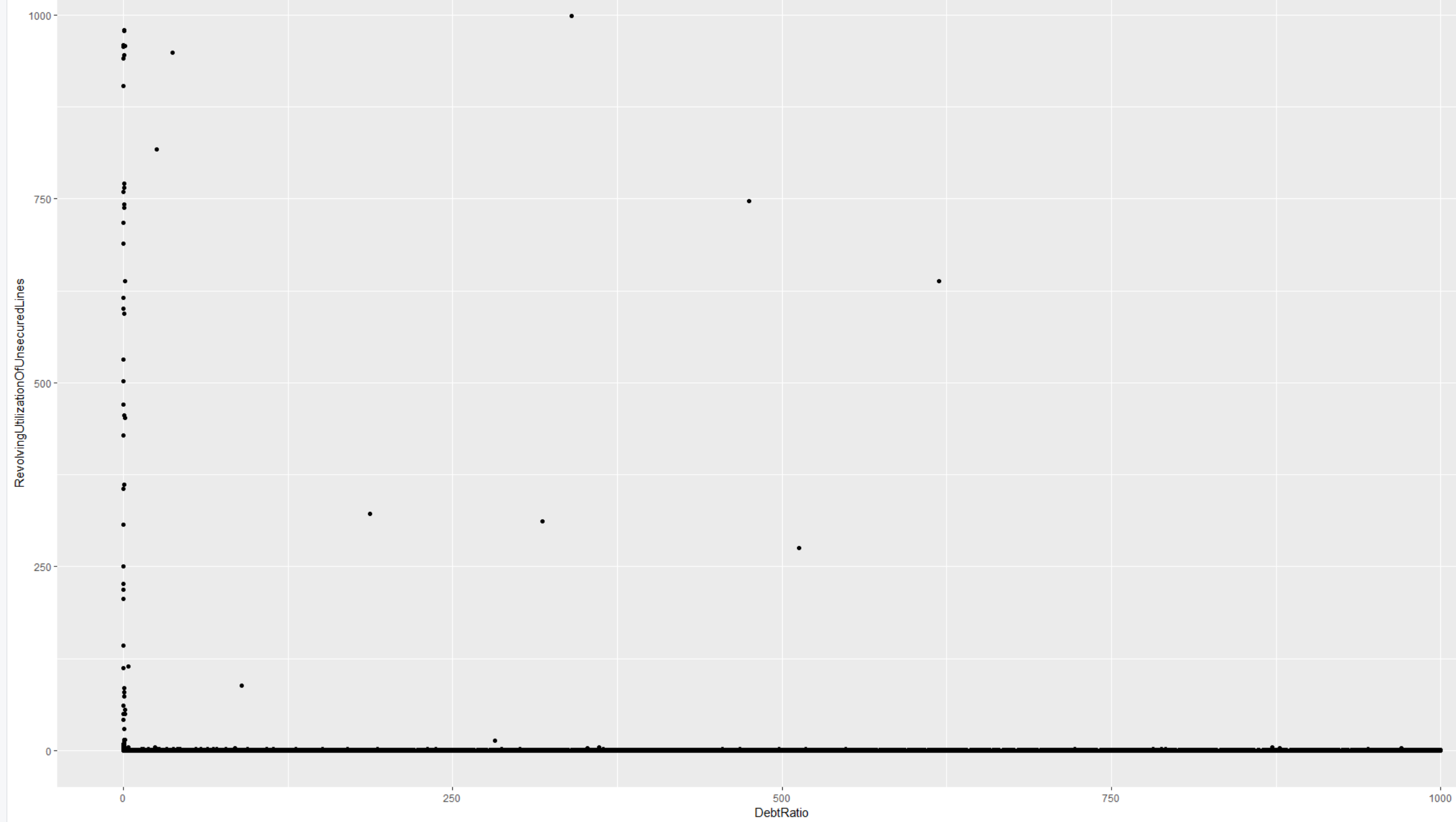
There is a very wide range of monthly income for lesser number of open credit lines and loans. As the number increases a negative correlation could be observed.

1. (x=age,y=MonthlyIncome)



More number of people between the age of 40-65 have higher and wider range of the monthly income.

1. (x=DebtRatio,y=RevolvingUtilizationOfUnsecuredLines)



Revolving utilization of unsecured lines are concentrated at low debt ratio and is constant at higher debt ratio throughout.

**Hypothesis Testing:**

1. T-test:

SeriousDlqin2yrs vs all the variables.

* MonthlyIncome

The p-value is 2.2e-16. Therefore null hypothesis is rejected and difference in means of monthly income of both defaulters and non-defaulters is not equal to 0.

* RevolvingUtilizationofsecuredlines

The p-value is 0.2248. Therefore null hypothesis is failed to be rejected and there is no significant difference in means of RevolvingUtilizationofsecuredlines of both defaulters and non-defaulters.

* NumberofCreditLinesAndLoans

The p-value is 2.2e-16. Therefore null hypothesis is rejected and difference in means of NumberofCreditLinesAndLoans of both defaulters and non-defaulters is not equal to 0.

* Debtratio:

The p-value is 4.835e-06. Therefore null hypothesis is rejected and difference in means of debt ratio of both defaulters and non-defaulters is not equal to 0.

* NumberOfTime30.59DaysPastDueNotWorse:

The p-value is 2.2e-16. Therefore null hypothesis is rejected and difference in means of debt ratio of both defaulters and non-defaulters is not equal to 0.

* NumberOfTime60.89DaysPastDueNotWorse:

The p-value is 2.2e-16. Therefore null hypothesis is rejected and difference in means of NumberOfTime60.89DaysPastDueNotWorse of both defaulters and non-defaulters is not equal to 0.

* NumberofRealEstateLoansOrLines:

The p-value is 0.02836. Therefore null hypothesis is rejected and difference in means of NumberofRealEstateLoansOrLines of both defaulters and non-defaulters is not equal to 0.

* Number of dependents:

The p-value is 2.2e-16. Therefore null hypothesis is rejected and difference in means of Number of dependents of both defaulters and non-defaulters is not equal to 0.

* NumberofTimes90DaysLate:

The p-value is 2.2e-16. Therefore null hypothesis is rejected and difference in means of NumberOfTime90DaysPastDueNotWorse of both defaulters and non-defaulters is not equal to 0.

1. One-way AOV test:

* RevolvingUtilizationofsecuredlines & SeriousDlqin2yrs

The p-value is 0.485. There is no statistical significant difference in the means of both groups(defaulters & non-defaulters).

* DebtRatio & SeriousDlqin2yrs

The p-value is 0.00324.There is statistical significant difference in the means of both groups.

* MonthlyIncome & SeriousDlqin2yrs

The p-value is 3.11e-12. There is statistical significant difference in the means of both groups.

1. Two-way AOV test:
2. Monthly Income vs Age\_group and SeriousDlqin2years

The population means of the individual groups are not equal but the combinations are significantly equal.

1. Debt ratio vs Age\_group and SeriousDlqin2year

The population means of the individual groups are not equal but the combinations are equal.

1. RevolvingUtilizationofUnsecuredLines vs Age\_group and SeriousDlqin2year

The population means of the age groups are not equal but population mean of SeriousDlqin2year and the combinations are equal.

1. NumberofCreditLinesAndLoans vs Age\_group and SeriousDlqin2year

None of the group’s population means are equal

1. NumberofTimes90DaysLate vs Age\_group and SeriousDlqin2year

None of the group’s population means are equal

1. NumberOfTime60.89DaysPastDueNotWorse vs Age\_group and SeriousDlqin2year

None of the group’s population means are equal

1. NumberOfTime30.59DaysPastDueNotWorse vs Age\_group and SeriousDlqin2year

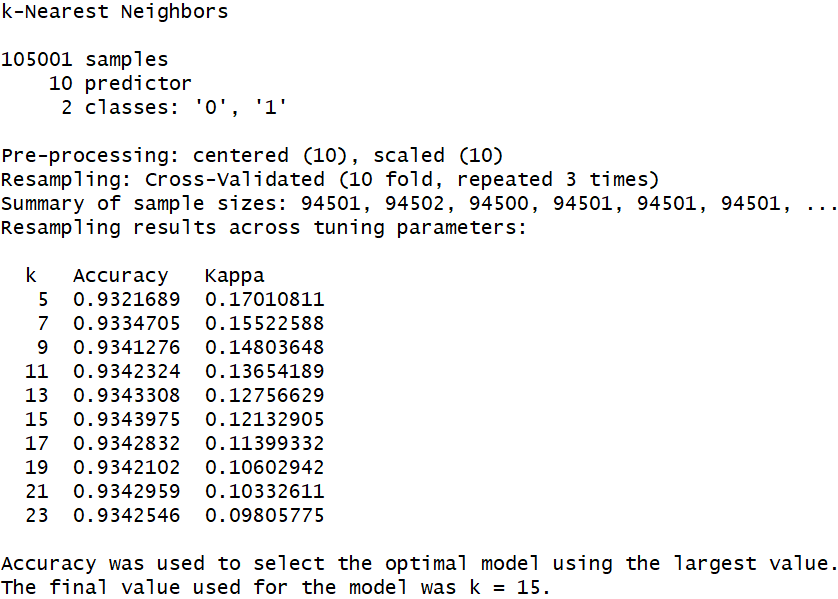
None of the group’s population means are equal

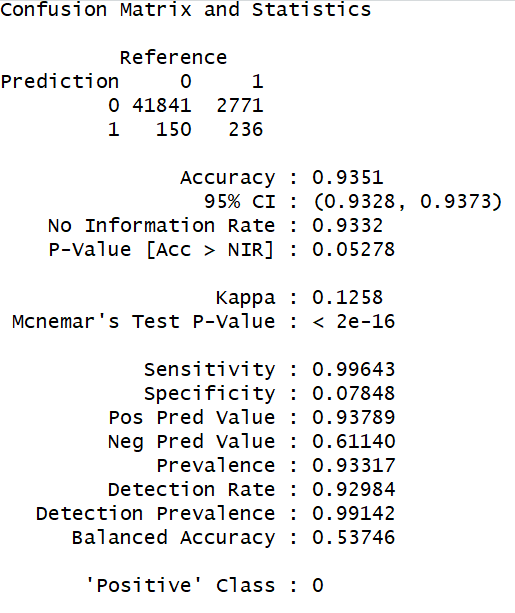
1. NumberofDependents vs Age\_group and SeriousDlqin2year

None of the group’s population means are equal

**Predictive Models:**

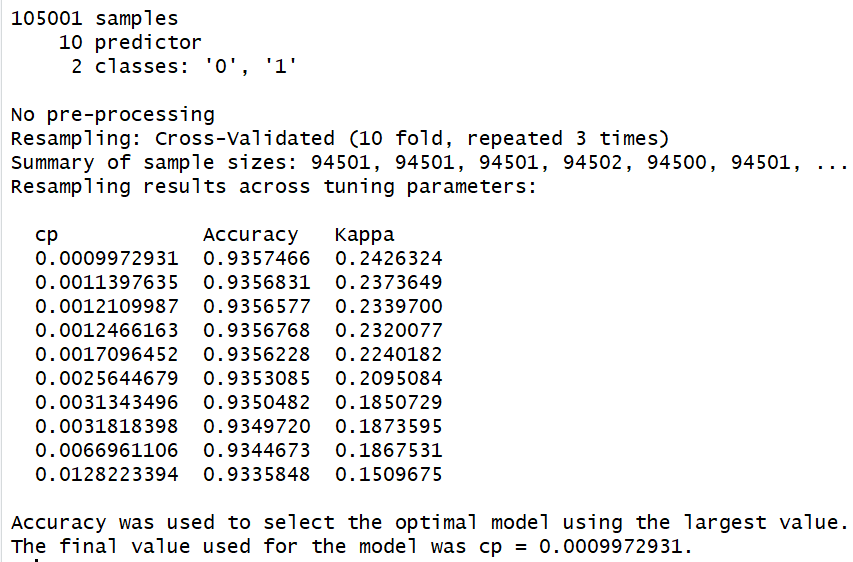
1. KNN

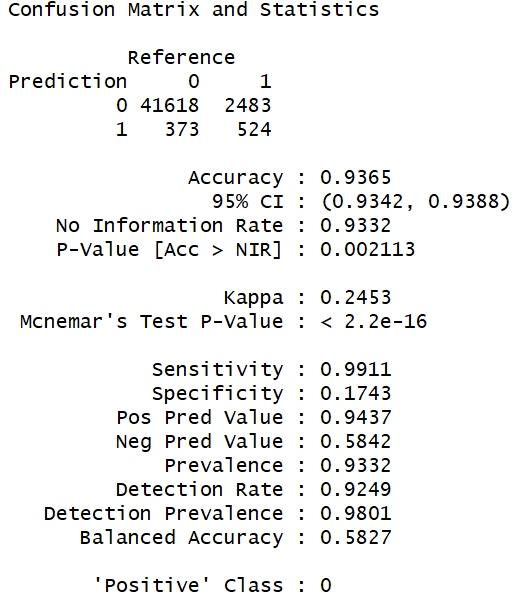




The KNN model was carried out between SeriousDlqin2yrs and all other variables except age\_group.K is determined as 15. The sensitivity is very high whereas the specificity is very low i.e this model is more accurate for non-defaulters(SeriousDlqin2yrs=0) than defaulters(SeriousDlqin2yrs=1). The accuracy for this model is 0.9351 with 95% confidence interval.

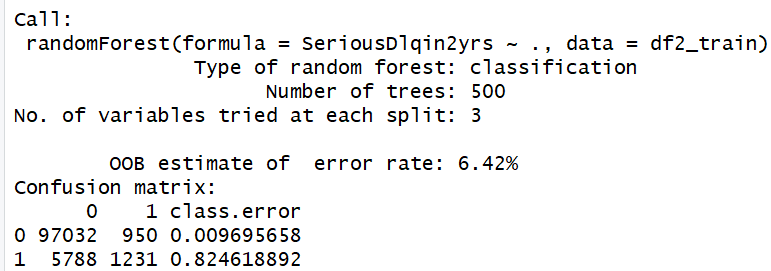
1. Decision Tree

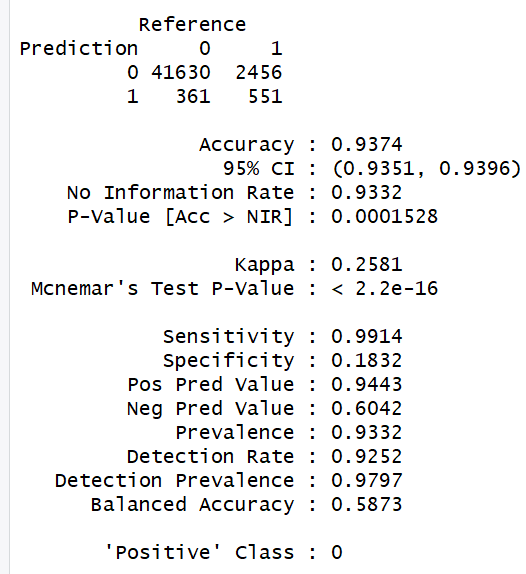




The decision tree was carried out between SeriousDlqin2yrs and all other variables except age\_group. The optimal complexity parameter for the decision tree is 0.0009972931. The accuracy of this model is 0.9365 with 95%confidence interval. The sensitivity is very high whereas the specificity is very low i.e this model is more accurate for non-defaulters (SeriousDlqin2yrs=0) than defaulters (SeriousDlqin2yrs=1).

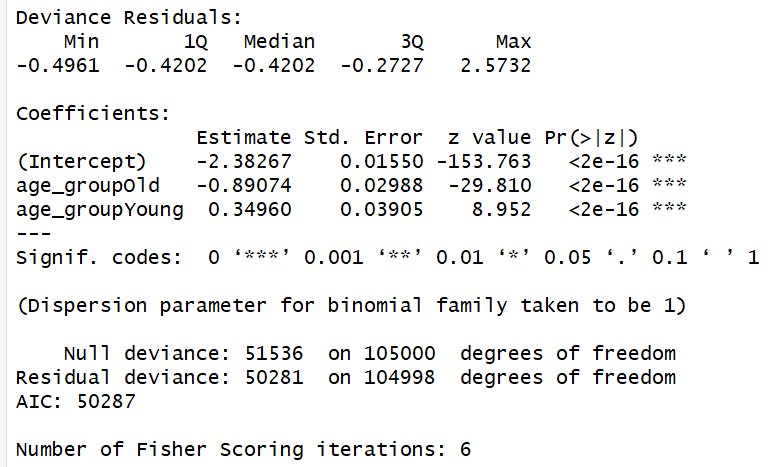
1. Random Forest

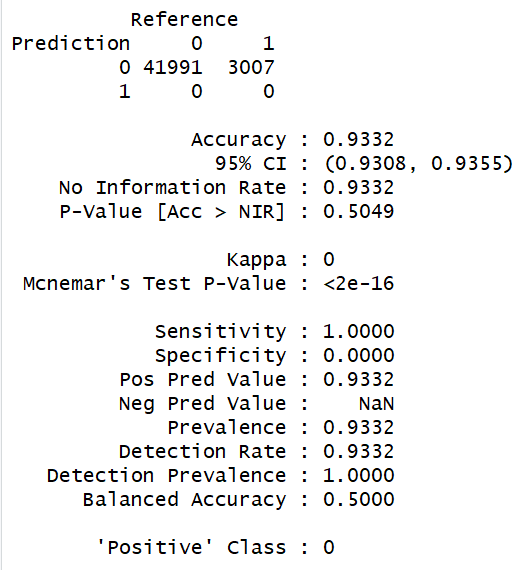




The random forest was carried out between SeriousDlqin2yrs and all other variables. The number of variables tried at each split is 3. The accuracy of this model is 0.9374 with 95%confidence interval. The sensitivity is very high whereas the specificity is very low i.e this model is more accurate for non-defaulters (SeriousDlqin2yrs=0) than defaulters (SeriousDlqin2yrs=1).

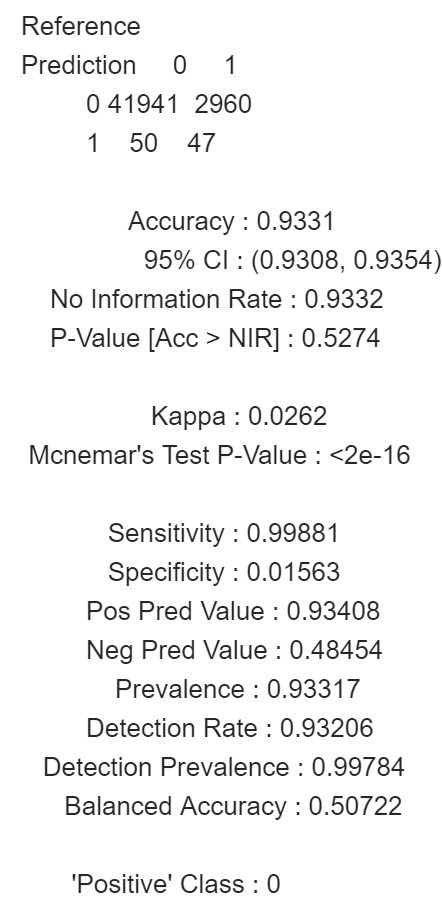
1. Logistic regression





A logistic regression was carried out between SeriousDlqin2yrs and age\_group. The accuracy is 0.9332. The sensitivity is 1 whereas the specificity is 0 i.e this model is accurate for non-defaulters (SeriousDlqin2yrs=0) than defaulters (SeriousDlqin2yrs=1).

1. SVMlinear



The accuracy of SVM linear model is 0.9331 and it has a high sensitivity and low specificity.

Key points:

* Old and middle age group people have almost the same range of real estate loans and lines and young people the least.
* The number of real estate loans and lines are of the same range for both defaulters and non-defaulters.
* The median and the range of RevolvingUtilizationOfUnsecuredLine is higher for young people.
* The debt ratio median is the highest for middle age and lowest for youngsters.
* The median debt ratio for defaulters is higher than non-defaulters.
* The monthly income range of both middleage and youngsters are almost the same and least for old people.
* The monthly income of non-defaulters have a larger range.
* The frequency of number of times the payment was exceeded 30-59 times significantly is at 96-98 and 7-13.
* The frequency of number of times the payment was exceeded 60-89 times significantly is at 6-11 and 96-98.
* The number of dependents are much higher for non-defaulters than defaulters.
* There is a very wide range of debt ratio and monthly income for low income and as income decreases debt ratio also decreases.
* The revolving utilization of unsecured lines are more concentrated at low income.
* Revolving utilization of unsecured lines are concentrated at lower debt ratio and is constant at higher debt ratio throughout.
* There is a very wide range of monthly income for lesser number of open credit lines and loans. As the number increases a negative correlation could be observed.
* More number of people between the age of 40-65 have higher and wider range of the monthly income.
* There is no significant difference in means of RevolvingUtilizationofsecuredlines of both defaulters and non-defaulters as observed in t-test.
* Accuracy of random forest model is the highest and is more apt for non-defaulters since the sensitivity is high.
* logistic regression was carried out between SeriousDlqin2yrs and age\_group. The accuracy is 0.9332 and this model is accurate for only non-defaulters since sensitivity is 1.
* The median and the range of RevolvingUtilizationOfUnsecuredLine is higher for defaulters.
* There is a very wide range of debt ratio and monthly income for low income and as income decreases debt ratio also decreases.