**Data set introduction:**

Banks play a crucial role in market economies. They decide who can get finance and on what terms and can make or break investment decisions. For markets and society to function, individuals and companies need access to credit.

Credit scoring algorithms, which make a guess at the probability of default, are the method banks use to determine whether or not a loan should be granted. This competition requires participants to improve on the state of the art in credit scoring, by predicting the probability that somebody will experience financial distress in the next two years.

**Data dictionary**

**SeriousDlqin2yrs** Person experienced 90 days past due delinquency or worse

**RevolvingUtilizationOfUnsecuredLines**: Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits

**age** Age of borrower in years

**NumberOfTime30-59DaysPastDueNotWorse**: Number of times borrower has been 30-59 days past due but no worse in the last 2 years.

**DebtRatio**: Monthly debt payments, alimony,living costs divided by monthy gross income

**MonthlyIncome**: Monthly income

**NumberOfOpenCreditLinesAndLoans**: Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)

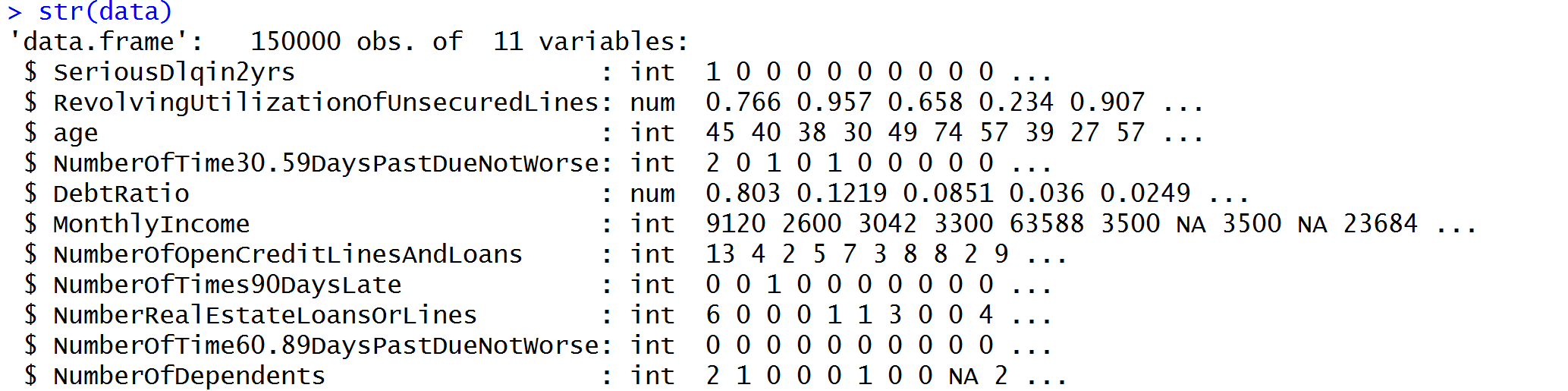
**NumberOfTimes90DaysLate**: Number of times borrower has been 90 days or more past due.

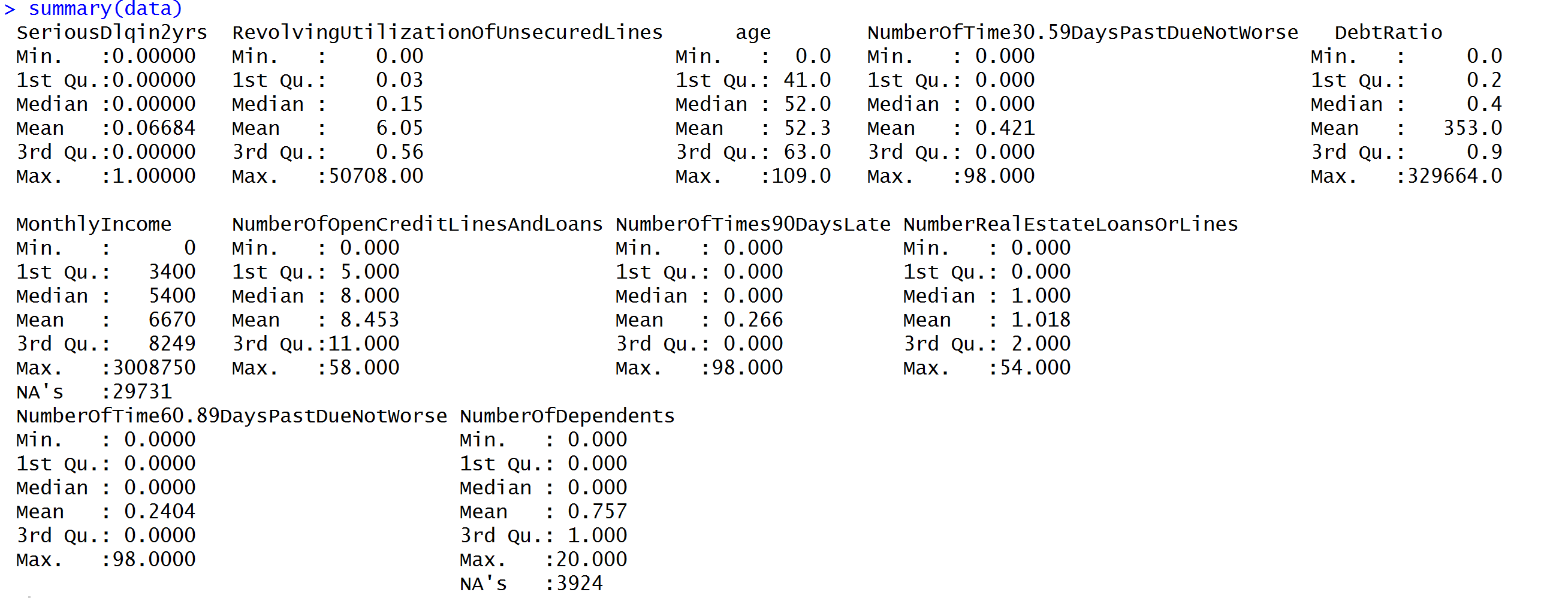
**NumberRealEstateLoansOrLines**: Number of mortgage and real estate loans including home equity lines of credit

**NumberOfTime60-89DaysPastDueNotWorse**: Number of times borrower has been 60-89 days past due but no worse in the last 2 years.

**NumberOfDependents**: Number of dependents in family excluding themselves (spouse, children etc.)

The goal of this competition is to build a model that borrowers can use to help make the best financial decisions.

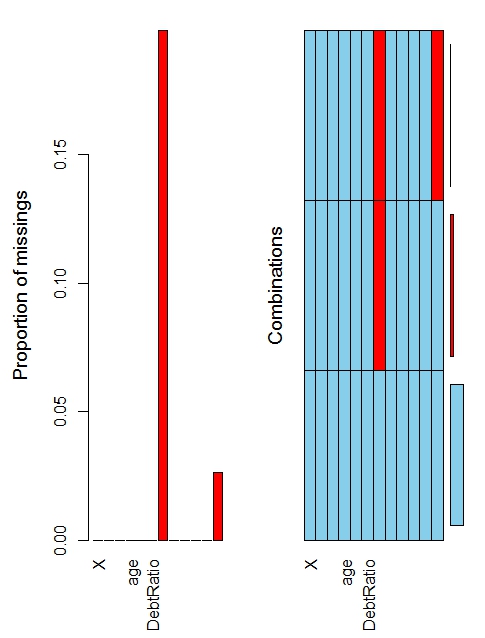


****

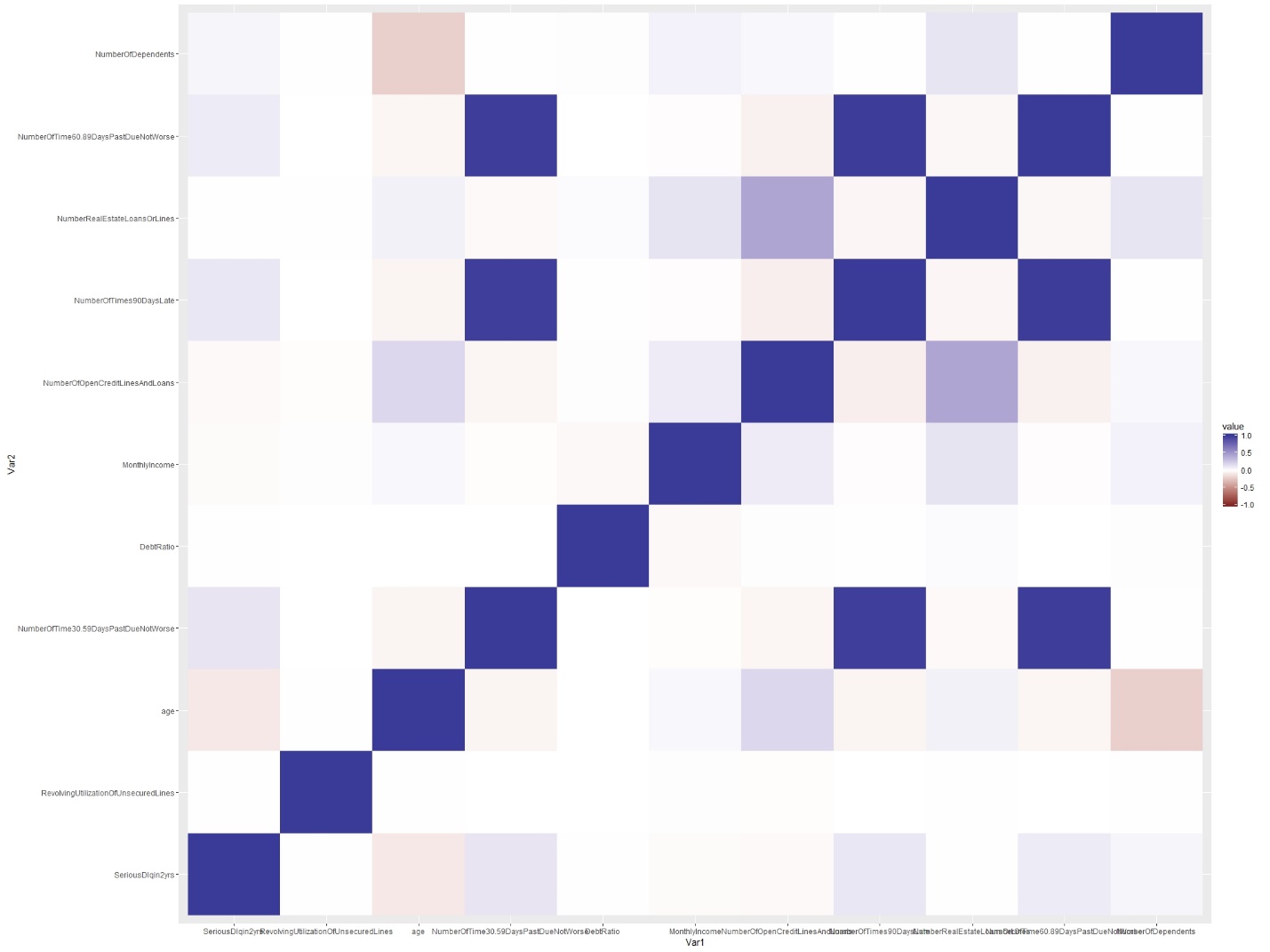
**EDA:**

**Mssing Value:**

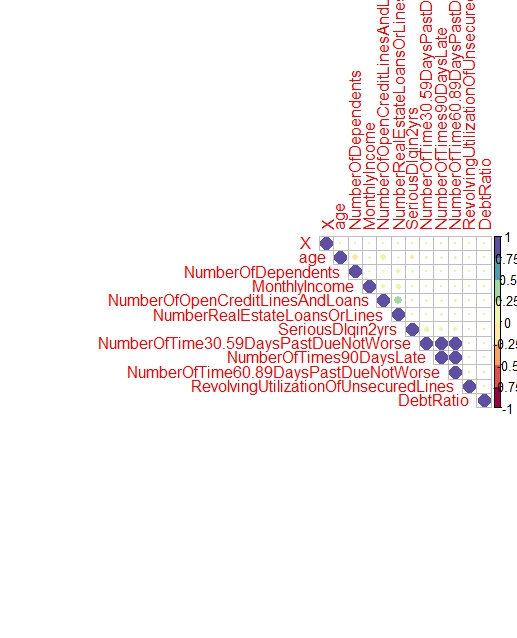
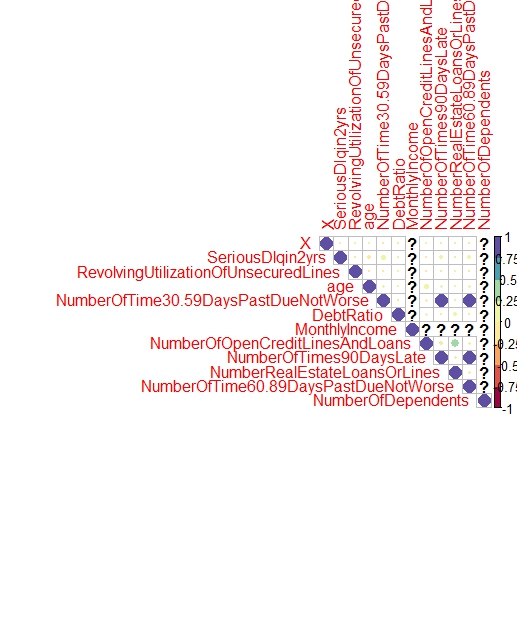
From the above data summary, we can see that variable monthly income has 29731 NAs (around 19.8% missing rate) and NumberofDependents has 3924 NA’s (2.6%). However, there are weird numbers in column debtratio and RevolvingUtilizationOfUnsecuredLines columns. Need to deal with that.

****

**Correlation Matrix:**

Below is the correlation matrix heat map. It showed that variables NumberOfTime30-59DaysPastDueNotWorse, NumberOfTime60-89DaysPastDueNotWorse, NumberOfTime90Dayslate are highly correlated with each other. In fact, the correlation is almost 1.0. 

These graphs also showed the same thing:



**PCA:**

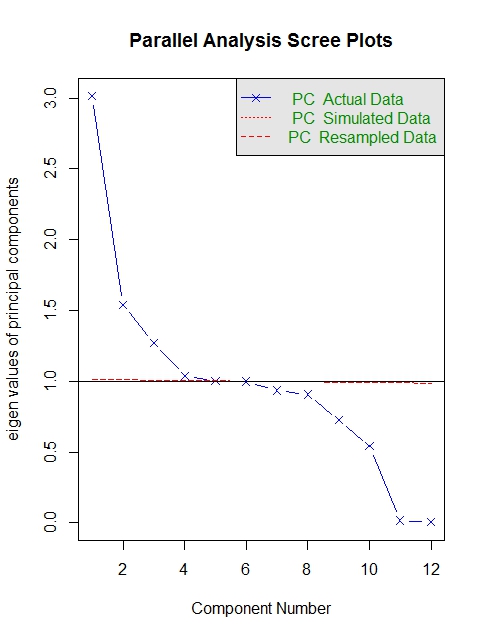
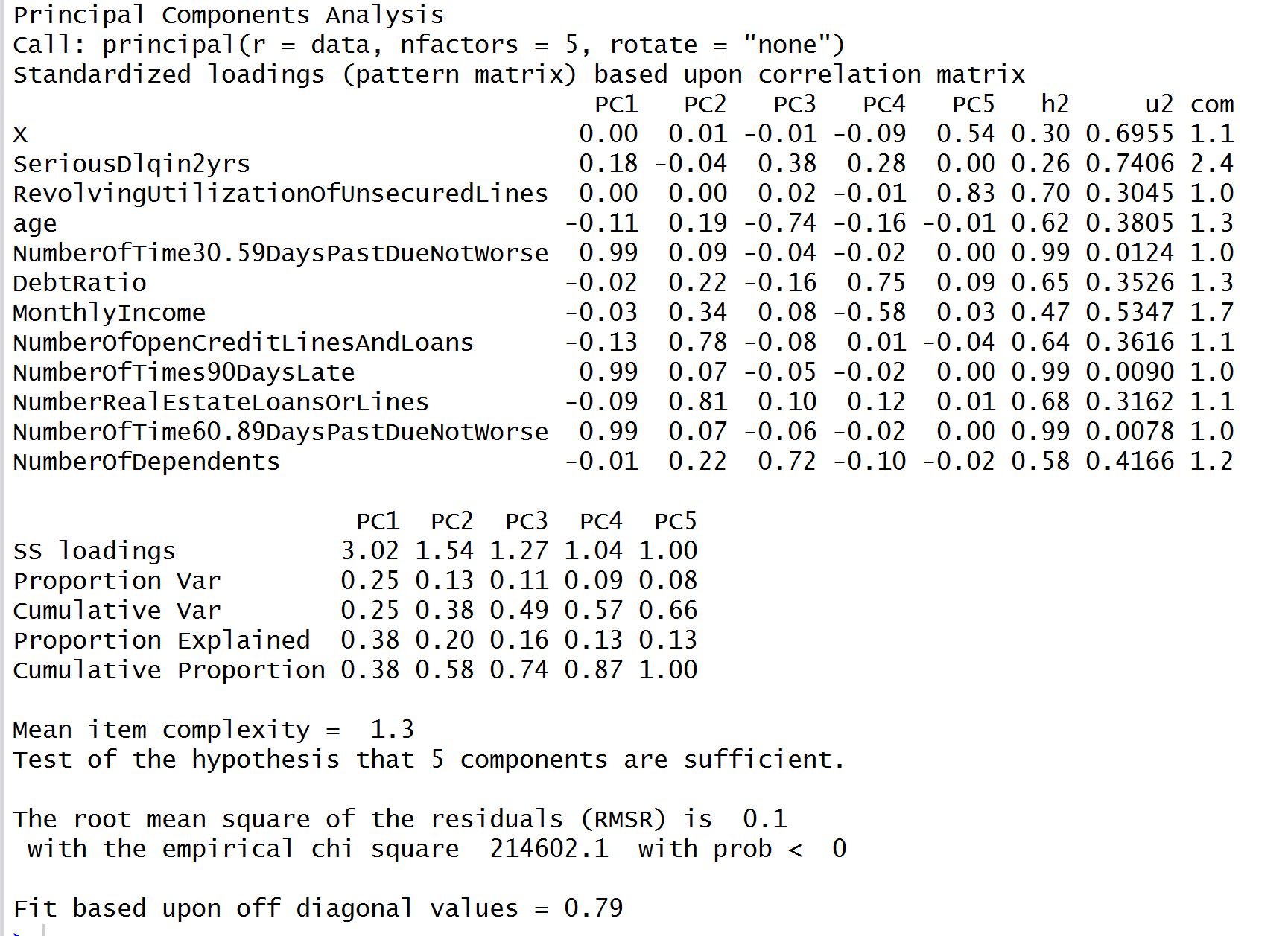
PC1: Numberof – daysPastDue

PC2: NumberOfOpenCreditLinesAndLoans, NumberRealEstateLoansOrLines

PC3: Age, NumberofDependents

PC4: DebtRatio

PC5: RevolvingUtilizationOfUnsecuredLines

****