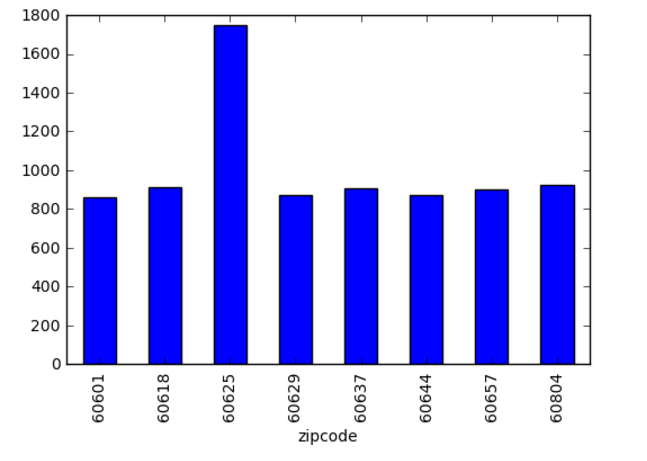
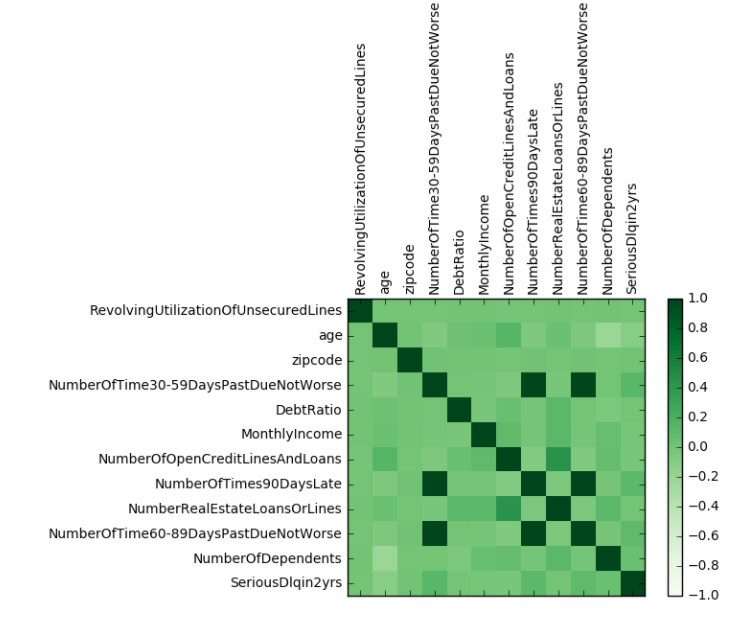
**Can we predict if a person experienced 90 days past due delinquency or worse?**

For this report we took the credit data for all people and need to find the probability that a person is a serious delinquent. Predictably, delinquency is linked with younger age, lesser income, lower debt ratio and more number of dependents. Hence the highest number of people experienced 90 days past due delinquency or worse was in the zipcode 60625



Apart from this I checked if there is any correlation between the predictor variables and the predicted variable. Surprisingly, there was no high correlation between the variables.



Then I performed data cleaning for the data as many of the variables (Monthly income and Number of dependents) had no values. Before updating the data, I had split up the data as testing and training to avoid knowledge spillover. Updated the income with the median income as the variance of income is higher. So median is a better measure as compared to mean as it is less sensitive. Similarly, I updated number of dependents with the median value of 0.To generate features, I divided the age into 10 bins and converted them into binary values. I performed a similar function for zip codes. Since both the incomes and debt\_ratio have high variance, I performed scalar transformation for them.

Finally I build a logistic regressor classifier for our training data and tested it out on our test set.

I built two models, once for all of our features and the second time with only our best 10 features after removing correlated features ie, the age bins. Surprisingly, both the models performed similarly. To test our model, I constructed a base model which assumed a prediction of zero for all predicted variable irrespective of input parameter variables. This performed as well as our model!

Hence I concluded that our model will not be able to predict our output ie, whether a person experienced 90 days past due delinquency or worse.