Introduction

The purpose of this paper is to explain the level of interest rates for the loans (over an unspecified time frame) related to the activity of the *Lending Club*[1].

The data supplied have been uploaded, modified and analysed using the R software (with few additional packages on top of the base ones[2]).

A statistical model is proposed in order to explain the level of the interest rate; the same model is also tested in terms of predictive power by dividing the given dataset between a training and a test set.

The main conclusions from the study is that the *Lending Club* setting of interest rates is less impacted than expected by some fundamental variables (balance to income, debt to income, length of the employment); the relationship with other variables (FICO rate[3], amount, length, inquiries, open lines and monthly income) appears to be significant and with the expected sign.

The vast majority of the method and the work flow used has heavily borrowed from the three Coursera[4] courses the author attended this year: *Introduction to data science* (prof. Bill Howe – University of Washington)*, Computing for data analysis* (prof. Roger Peng – Johns Hopkins Bloomberg school of public health) and *Data analyisi* (prof. Jeff Leek – Johns Hopkins Bloomberg school of public health).

All mistakes are obviously mine. The author is: *matteo.castagna@gmail.com*

Methods

Data collection and variables management

The initial work consisted in loading the data, replacing the proposed variable names to a more manageable ones and creating extra variables using the ones supplied.

The initial dataset has been uploaded from a CSV file (downloaded on *Thu Nov 14 22:41:06 2013*) with the initial option set to *as.is = TRUE*. That has required the coercion to numerical for variables initially loaded as characters (*Debt to income, Interest rate, Employment length*) and addition of factors for others (*Loan Length, Home Ownership, Loan Purpose, State*).

The manipulation of the *char* FICO rate ranges consisted in:

- creating a factor based on the range
- adding a variable with the centre of each the range as numerical value.

Three more variable has been added and used on top of this:

Balance To Income: $\frac{\textit{Revolving Balance}}{\textit{Monthly Income}}$ as an indicator of the potential solvency of the borrower

Funded Percentage: $\frac{\textit{Amount funded}}{\textit{Amount requested}}$ as an indicator of the workings of the *Lending Club*

Loan Purpose Count: number of occurrences in the sample of the relevant Loan purpose. This has been used to filter out the instances of *Purpose of the loan* with a lower number of cases when creating the separate picture.

Exploratory analysis

Few odd values has been replaced with 0 (*Loan Amount* = -0.01)[5] or NA (Home ownership = NONE).

The number of complete cases (observations without any NAs for all the 24 variables created – 10 more than the 14 provided) is equal to 2421. 79 observations have an NA in one or more variables; effectively all but two are related to the newly created numerical *Employment length* variable.

For statistical modelling purpose only the complete cases have been used. The loss of information from eliminating 79 partial observations is effectively not relevant.

Some thought has been given on how to treat the Loan Amount = -0.01 but at the end, because of how the *Lending Club* works, it was decided that this was equivalent to a "loan offer not taken" and then considered as a zero value.

This is debatable but, ultimately, not very important for the purpose of the analysis.

The exploratory analysis immediately made clear how important is the FICO rate in order to determine the level of interest rate of the deal (Illustration 1).

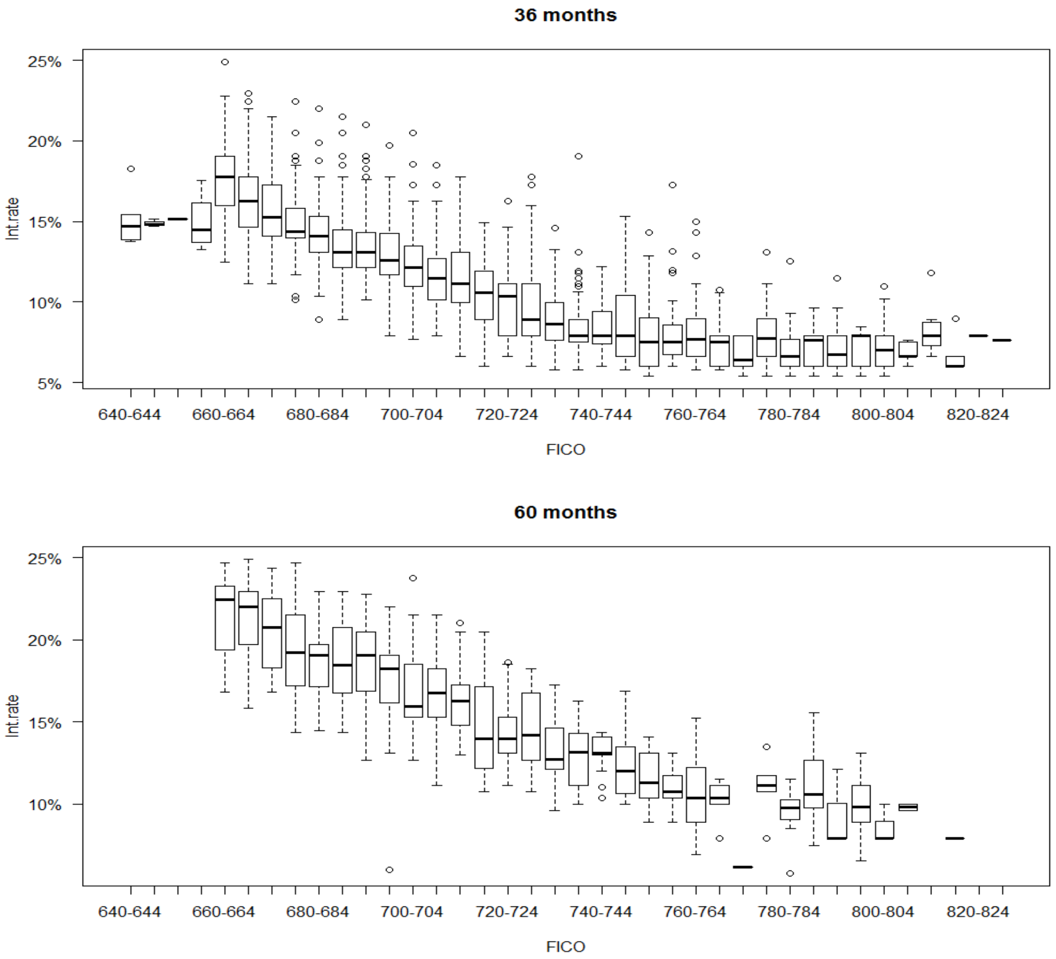

Illustration 1: Interest rate vs. FICO rate

At the same time it was clear that the FICO rate link to the variable $\frac{\textit{Amount conceded}}{\textit{Amount requested}}$ wasn't particularly interesting beyond very low level of that rate (see Illustration 3).

In order to exclude the presence of *confounders* a basic correlation analysis among the possible dependent variables has been run.

Illustration 2 presents the basic results hinting at an effective orthogonality among the different variables[6].

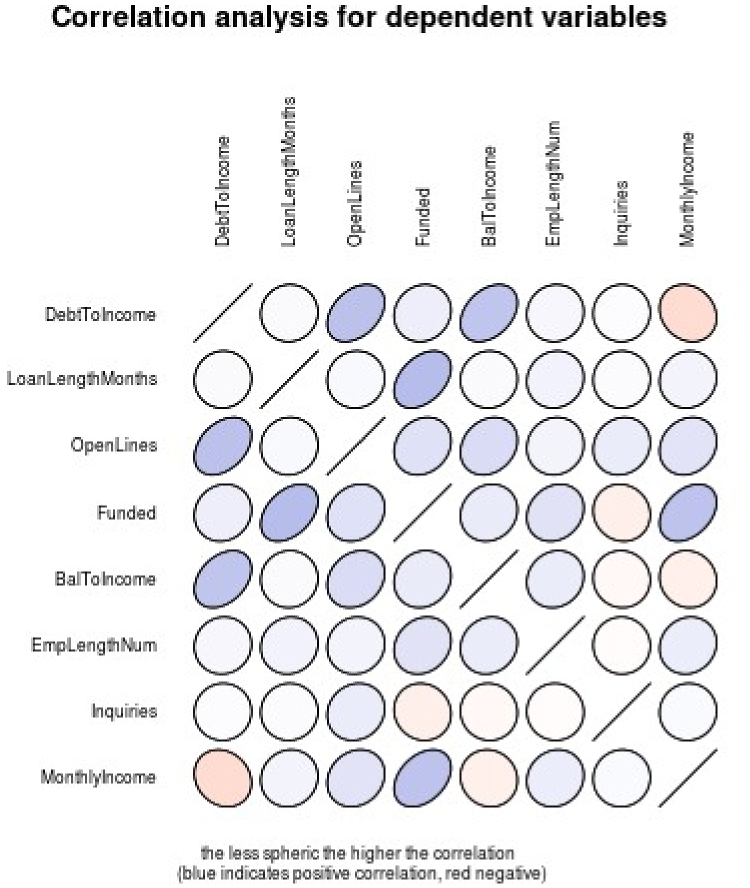

Illustration 2: Correlation between dependent variables

You can appreciate a mild correlation for the pairs *DebtToIncome-OpenLines, DebtToIncome-BalanceToIncome, Funded-MonthlyIncome, LoanLengthMonths-Funded.*

The correlation coefficients is fairly low at around 0.4 at best in these cases, suggesting that all of these could be used as independent explanatory variables.

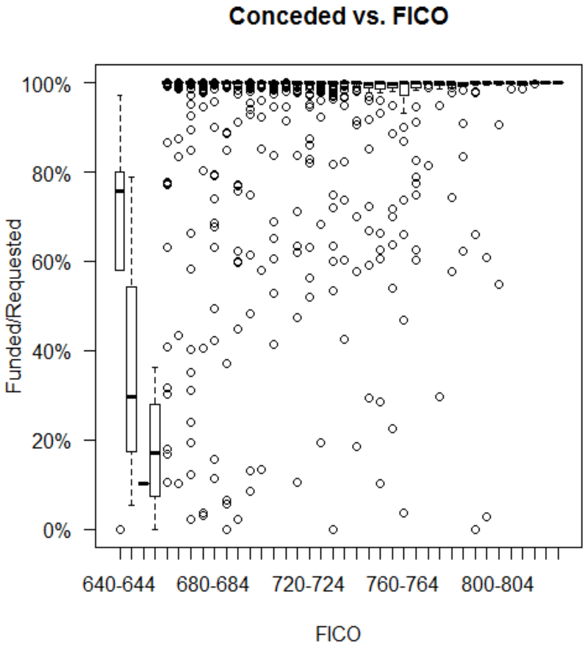

Illustration 3: Amount funded/Requested

On a more interesting note the exploratory analysis hinted to a strong effect of the duration of the lending: the longer maturity ones have substantially higher interest rates (see illustration 4).

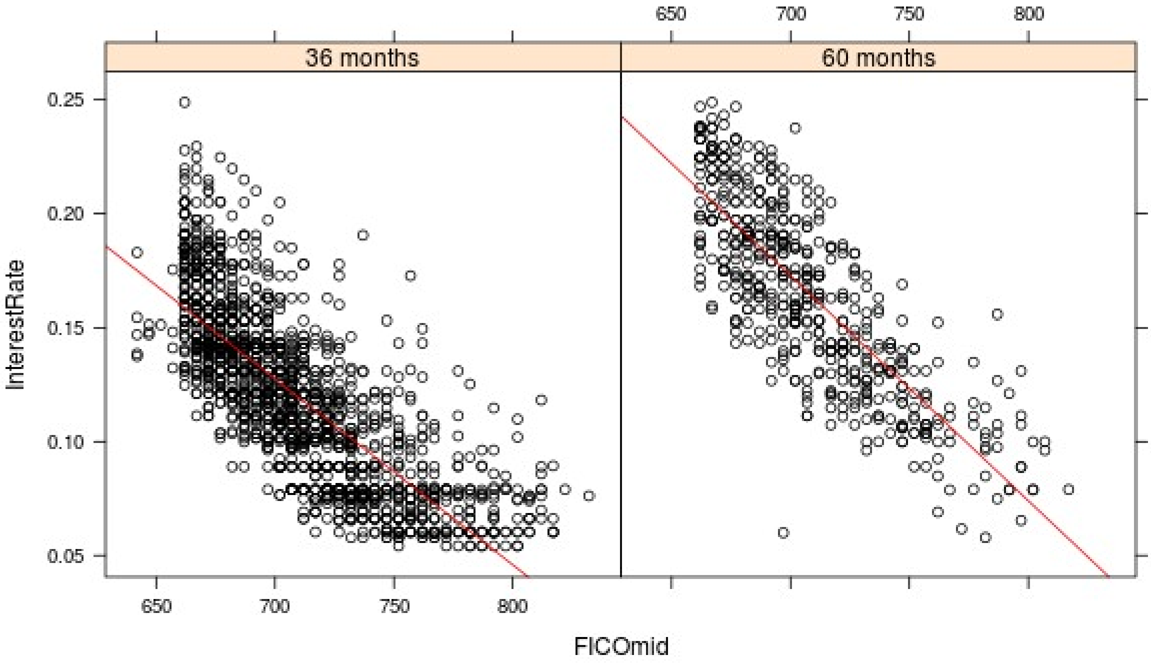

Illustration 4: Interest rate vs. FICO by maturity

The low relevance of some standard fundamental variables can be exemplified by illustration 5 which clearly hints to an absence of relationship between the interest rate of the loan and the *debt to income* level of the borrower.

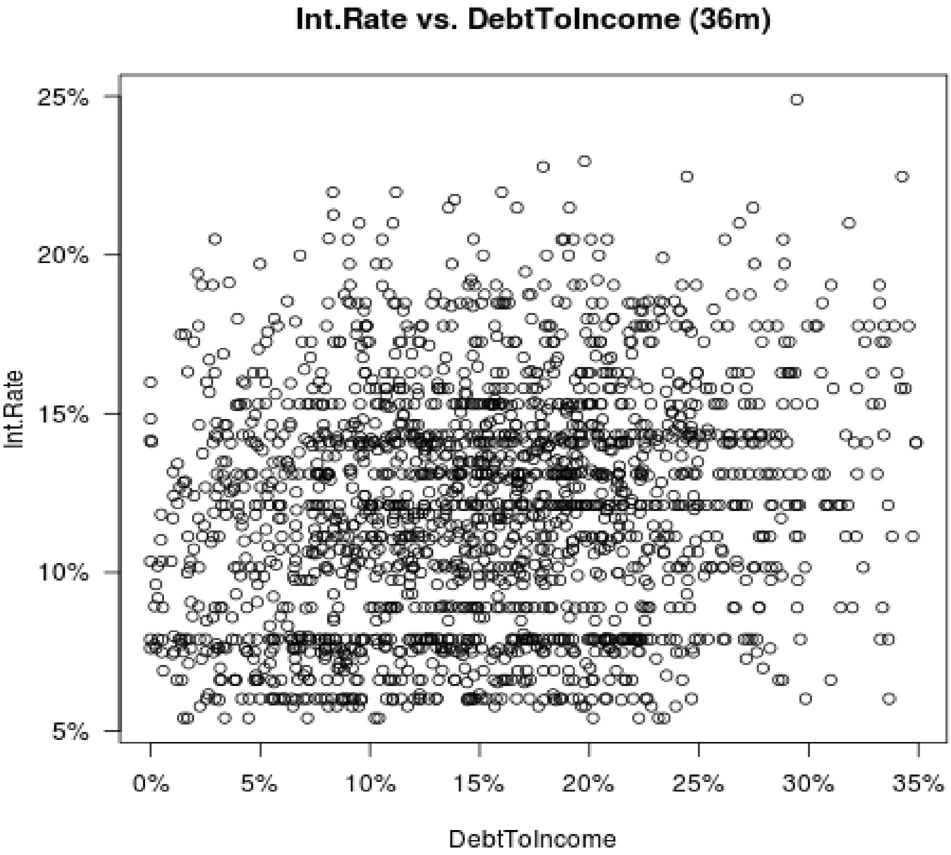

Illustration 5: Interest rate (loans of 36 months) vs. Debt to Income ratio

Statistical modelling

In order to model the relationship among the different variables in the data set a standard multivariate linear regression model (the *lm()* function in R) has been used and a discretionary approach in order to achieve the minimal adequate model[7] was applied.

Results

The results proposed by our preferred statistical model are as follows:

no scaling

```
Residuals:
      Min        1Q    Median        3Q       Max 
-0.094867 -0.013553 -0.001813  0.012002  0.098433 
```

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            7.186e-01  8.685e-03  82.744  < 2e-16
FICOmid               -8.637e-04  1.212e-05 -71.259  < 2e-16
LoanLengthFac60 months  3.207e-02  1.101e-03  29.133  < 2e-16
OpenLines             -4.244e-04  9.565e-05  -4.437 9.53e-06
Funded                 1.545e-06  6.450e-08  23.953  < 2e-16
Inquiries              3.881e-03  3.425e-04  11.333  < 2e-16
MonthlyIncome         -2.627e-07  1.141e-07  -2.302   0.0214

Residual standard error: 0.02044 on 2414 degrees of freedom
Multiple R-squared:  0.7589,	Adjusted R-squared:  0.7583 
F-statistic:  1266 on 6 and 2414 DF,  p-value: < 2.2e-16
```

confidence interval for the estimates

```
                               2.5 %        97.5 %
(Intercept)            7.016085e-01  7.356706e-01
FICOmid               -8.874218e-04 -8.398884e-04
LoanLengthFac60 months  2.991244e-02  3.422992e-02
OpenLines             -6.119325e-04 -2.368219e-04
Funded                 1.418431e-06  1.671381e-06
Inquiries              3.209811e-03  4.553074e-03
MonthlyIncome         -4.865240e-07 -3.888208e-08
```

zero does not belong to any of them confirming the previous results.

scaling (i.e. Beta analysis)

```
Residuals:
    Min      1Q  Median      3Q     Max 
-2.2812 -0.3259 -0.0436  0.2886  2.3670 

Coefficients:
                                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                                         -1.535e-15  9.992e-03   0.000   1.0000
scale(FICOmid)                                      -7.257e-01  1.018e-02 -71.259  < 2e-16
scale(as.numeric(sub(" months", "", LoanLengthFac)))  3.205e-01  1.100e-02  29.133  < 2e-16
scale(OpenLines)                                    -4.600e-02  1.037e-02  -4.437 9.53e-06
scale(Funded)                                        2.869e-01  1.198e-02  23.953  < 2e-16
scale(Inquiries)                                     1.153e-01  1.018e-02  11.333  < 2e-16
scale(MonthlyIncome)                                -2.519e-02  1.095e-02  -2.302   0.0214

Residual standard error: 0.4916 on 2414 degrees of freedom
Multiple R-squared:  0.7589,	Adjusted R-squared:  0.7583 
F-statistic:  1266 on 6 and 2414 DF,  p-value: < 2.2e-16
```

***the effect of the* loan purpose factor**

As presented by the separate picture[8], the effect of the Loan purpose factor looks relevant: the regression lines have different angle and intercept depending on the reasons for the loan. Statistically this can be explored by using the analysis of variance method.

```
Analysis of Variance Table
Response: InterestRate
                  Df Sum Sq   Mean Sq F value    Pr(>F)
LoanPurposeFac    13 0.1586 0.0121990  7.2923 3.416e-14
Residuals       2407 4.0266 0.0016729
```

The loans purpose factor is statistically significant, correlating with the level of the interest rate. Adding it to preferred the regression model presented above

a) it is expensive (it adds 13 dummy variables)

b) does not lead to a substantial improvement of the model (the adjusted R^2 goes from 0.7583 to 0.7622).

predictive power of the model

The sample of the complete cases has been divided in a test set and a training set (50% of the observations each) and the estimates of the model parameters obtained using the training set were used to predict the interest rates for the test set (out of sample predictions).

Using a fairly standard technique based on RMSD[9] we can then assess the predictive power of the model: the normalized root-mean-squared-deviation we obtained was 10.7% (that is the model predicted interest rate was wrong on average by 10.7% of the true interest rate). In absolute terms that means the model is, on average, wrong by 207 basis points (that is 2.07% *per annum*).
The proportion variation explained in the outcome of the testing data[10] is equal to 76%.
Graphically the result of the prediction exercise on the test set is presented with illustration 6 (where the red line has a slope of 45deg.)

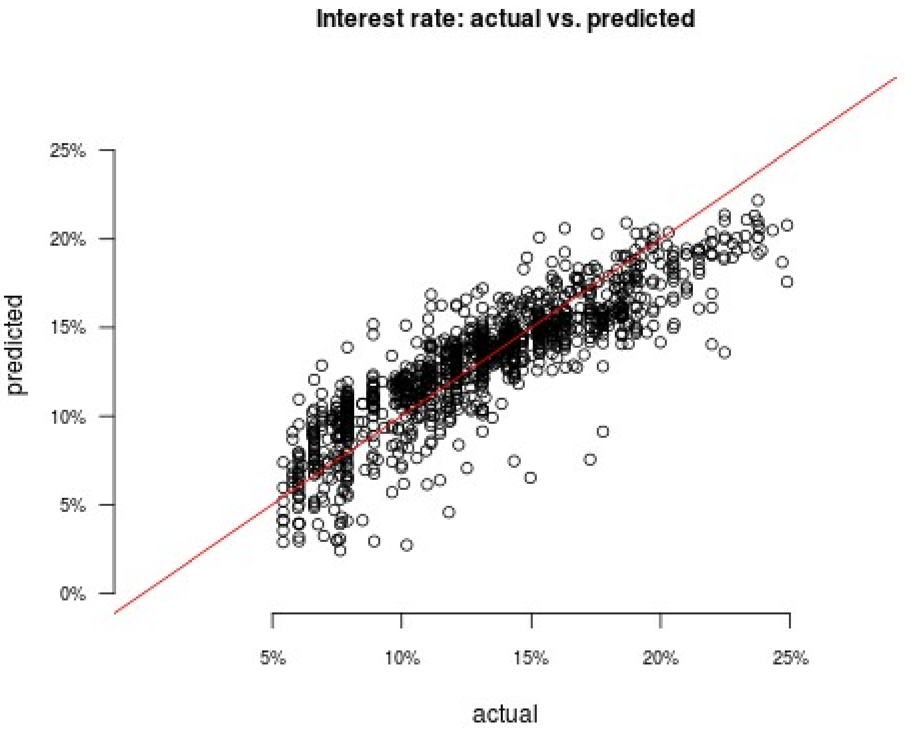

Illustration 6: actual vs. predicted interest rate (testing sample)

Conclusions

The FICO score provide the most important variable explaining the level of the interest rate as showed by the beta analysis in the scaled regression.

If the FICO score is worse by 1% the interest rate is higher by 0.73% or, using non-scaled variables, a FICO score 10 points higher decreases the level of interest rates by 86 basis points (0.86% per annum).

The length of the lending is relevant as well: 60 months loans are 320bps (3.2% per annum) higher than 32 months ones.

As expected the amount funded and the number of inquiries is positively related with the interest rate level: the higher the amount borrowed and the more difficult it was to get to it (as represented by the number of *Inquiries* over the last 6 months) the higher the interest rate.

The amount of open lines is negatively related to the interest rate: the more lines are open for an individual the lower the rate. Effectively this indicate that the solvency for that specific borrower was already successfully tested in the past and the lenders are more willing to part from their money. This interpretation can be challenged by saying that the more open lines a person has the more exposed the *loan club* is: this might lead to a weaker solvency profile.

A weakness of the data set is the absence of the time reference of those loans. We don't know anything about when the loans were made: this is relevant because the interest rate level (as represented by the Lending Club base rate) is a variable that should fluctuates over time. The market level of interest rates for the three and ten years period observed at the time of the deal could have been a logical explanatory variable.

As per the weaknesses of this study (little time the obvious excuse) it must be noted that only a very superficial look was given at the differences by State. It must be noted that it should be quite difficult to justify massive discrepancies in that dimension.

An other type of analysis not performed is the clustering of loans by purpose with a view on grouping the different cases and possibly improve the regression results with more dummy variables added to the final model.

Note as well that the results for loan maturity and amount of the loan are consistent with the Lending Club mechanics of setting the loan rate[11].

References and notes

[1] https://www.lendingclub.com/home.action

[2] Packages used: lattice, ggplot2, Rcurl, scales, ellipse

[3] FICO ranges from 300 to 850, the higher the better. Median for 2011 was 711 (2011). According to a Fitch study, the accuracy of FICO in predicting delinquency has diminished in recent years. In 2001 there was an average 31-point difference in the FICO score between borrowers who had defaulted and those who paid on time. By 2006 the difference was only 10 points." (http://en.wikipedia.org/wiki/Credit_score_in_the_United_States)

[4] http://www.coursera.com

[5] This is debatable but, ultimately, not very important for the purpose of the analysis.

[6] The numbers are as follows:

```
                 DebtToIncome LoanLengthMonths  OpenLines      Funded BalToIncome EmpLengthNum   Inquiries MonthlyIncome
DebtToIncome       1.00000000       0.02810365 0.37623582  0.09769687  0.35644541   0.04926690  0.01644445   -0.16421866
LoanLengthMonths   0.02810365       1.00000000 0.03575747  0.40355073  0.02669452   0.07593517  0.02040458    0.06881792
OpenLines          0.37623582       0.03575747 1.00000000  0.17922555  0.20934813   0.06576275  0.11458861    0.16424148
Funded             0.09769687       0.40355073 0.17922555  1.00000000  0.11650919   0.17059007 -0.07480782    0.36805239
BalToIncome        0.35644541       0.02669452 0.20934813  0.11650919  1.00000000   0.11318483 -0.03156981   -0.07015508
EmpLengthNum       0.04926690       0.07593517 0.06576275  0.17059007  0.11318483   1.00000000 -0.01345050    0.10535739
Inquiries          0.01644445       0.02040458 0.11458861 -0.07480782 -0.03156981  -0.01345050  1.00000000    0.03004329
MonthlyIncome     -0.16421866       0.06881792 0.16424148  0.36805239 -0.07015508   0.10535739  0.03004329    1.00000000
```

[7] http://ww2.coastal.edu/kingw/statistics/R-tutorials/multregr.html and http://ww2.coastal.edu/kingw/statistics/R-tutorials/simplelinear.html

[8] The separate picture was created using the ggplot2 package and using the techniques explained in http://www.cookbook-r.com/Graphs/

[9] http://en.wikipedia.org/wiki/Root-mean-square_deviation

[10]http://gettinggeneticsdone.blogspot.co.uk/2011/02/split-data-frame-into-testing-and.html

[11]https://www.lendingclub.com/public/how-we-set-interest-rates.action