**Title: Increase in FICO Credit Score is associated with a Decrease in Interest Rate for Peer-to-peer loans**

**Introduction: Peer-to-peer lending** is the practice where money is lent to various individuals, known as "peers", without having to go for a bank or using some other financial services. This lending takes place online on peer-to-peer lending companies' websites using various different lending platforms and credit checking tools. The most important variable to be considered while giving loan is the FICO score.
A FICO score is actually the **credit score in the United States,** which is a number representing the creditworthiness of a person, the likelihood that person will pay his or her debts.Widespread use of credit scores has made credit more widely available and less expensive for many consumers.

The duration for which the loan is taken is also taken into account while determining the interest rate. The length is usually categorized in two factors, which are 36 and 60.
Important variables which help in deciding the interest rate as shown by our analysis are Amount Requested and Amount Funded by Investors, these variables are utilized when we break the data by the FICO Range. These are most useful when we have same FICO scores in some cases.

## Methods:

### Data Collection:

For our analysis we have used the dataset provided by the coursera team on the course "Data Analysis" which is a data set consisting of **a sample of 2,500 peer-to-peer loans** issued through the Lending Club. The data was downloaded from https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv on November 14, 2013, using R programming language[3].

### Exploratory Analysis

Exploratory analysis was performed by examining tables and plots of the observed data. We identified correlation and relationships of significance between variables on the basis of plots and knowledge of the scale of measured variables. Exploratory analysis was used to (1) identify missing values, (2) verify the quality of the data, and (3) determine the terms used in the regression model relating Interest rate to FICO Values. And Also to variables like Loan Length , Amount Requested and Amount Funded by investors when FICO values are equal.

Statistical Modeling
To relate loan Interest rate to FICO values we performed a standard multivariate linear regression model [4]. Model selection was performed on the basis of our exploratory analysis and prior knowledge of the relationship between FICO scores and Interest rate and also the relation between FICO values with different other variables such as Amount Requested in the loan application, amount loaned to the individual, and duration for which the loan was given.

**Reproducibility**
All analyses performed in this manuscript are reproduced in the R markdown file LoansDataFinal.Rmd [6]. To reproduce the exact results presented in this manuscript, the code should be executed on the same data set, and since loans are being given on a regular basis, the data is subjected to change.

**Results:**

The loansData consist of the important information for determining the relationships between different variables. It covered the Interest Rate (in %) for lending, the length of loan lending period (in months), the Amount requested and given (in dollars) . These were the variables which mainly showed some interesting patterns for our analysis. For the rest of the variables no significant patterns were observed which could be interesting for our analysis.

Since the distribution of Interest Rates are follow Normal distribution quite closely, we don't need a transform for it. Neither is done for the FICO values although they are converted to lower and higher values in numeric data type rather than complete range.

We fit a linear regression model for Interest.Rate and FICO (The values broken down into numerics from range factors).

The residuals showed patterns of non-random variation. We attempted to explain those patterns by fitting models including potential confounders. Ourl regression model was:

$$IR = b0 + b1\ FICO + f(LL) + g(AR) + h(AFI) + e$$

where b0 is an intercept term and b1 represents the change in interest rate in percentage on the Richter scale associated with a change of 1 unit for FICO scores at the same period of loan length, amount requested, and amount funded by investors. The terms g(AR), and h(AFI) represent factor models with 4 different levels each for the amount requested and funded in dollars. The term f(LL) is a factor variable with 2 different levels for the period of loan in months. The error term e represents all sources of unmeasured and unmodeled random variation in interest rates. Our regression model appeared to remove most of the non-random patterns of variation in the residuals.

However, we noticed that there was not much change in residual plot if we removed Amount Requested variable from the model and that the P values of the same was in the range .1 to .2 which is not that significant to our analysis. So instead we can use it with just

IR = b0 + b1 FICO + f(LL) + h(AFI) + e

to conduct our analysis. where b0 is an intercept term and b1 represents the change in interest rate in percentage on the Richter scale associated with a change of 1 unit for FICO scores at the same period of loan length, amount requested, and amount funded by investors. The term h(AFI) represent factor model with 4 different levels each for the amount requested and funded in dollars. The term f(LL) is a factor variable with 2 different levels for the period of loan in months. The error term e represents all sources of unmeasured and unmodeled random variation in interest rates.

We observed a very significant statistical association between Interest rate and FICO scores (< 2e-16). A change of one unit of FICO score corresponds to change of -0.087 on Ritcher scale(95% Confidence -0.0892, -0.084). The minus sign shows our exploratory analysis that there's a negative association between them.

**Conclusions:**
Our analysis suggests that there is a significant, negative association between the credit score FICO (Obtained by breaking FICO.Range into numeric lower and higher range values) and Interest Rate. This makes sense, as the people would more credit score get to have low interest rates.

We also observed that other variables such as Loan.Length , Amount.Funded.By.Investors, Amount Requested are related with both Interest rates and FICO Scores. Including these variables in the regression model relating interest rate to FICO scores improves the model fit, but does not remove the significant positive relationship between the variables. However, upon seeing the less significant P-values for Amount.Requested and no change in residual plot on removing it from model, we concluded that we can just use Loan.Length and Amount.Funded.by.Investors(converted into factor of level-4) in our final model which improves the model fit in same way and does not remove the relationship.

While our analysis is of interest for this data set which is of limited samples of 2500 peer-to-peer loans issued through the Lending Club. A larger collection of peer-to-peer loans may be more appropriate for understanding the relationship between Interest Rates and some other significant variables, in general and in cases where the FICO score is same for two individuals.
Our analysis may be of interest to individuals who are planning to give peer loans in the future just to get an idea from the picture. But for corporations who are running this type of loan systems or are just starting with it, lack of data might not give a full picture.

**REFERENCES:**

1.http://en.wikipedia.org/wiki/Peer-to-peer_lending
2.http://en.wikipedia.org/wiki/Credit_score_in_the_United_States
3.R Core Team (2012). ”R: A language and environment for statistical computing.” URL: http://www.R-project.org
4.www.stat.yale.edu/Courses/1997-98/101/linmult.htm Multiple Linear Regression
5.stattrek.com/**regression**/**residual**-analysis.aspx
6.R Markdown Page. URL: http://www.rstudio.com/ide/docs/authoring/using_markdown.