# Loans Interest Analysis

This is a short Analysis of data involving peer-to-peer loans, as collected and presented by the Lending club in the US (see details in this LINK)
The lending club is approving or rejecting a loan request and determines the interest rate for the loan.

## The Research Question

What variables determine the interest rate determined by the club?

## Methodology

The statistics of human decision

Since this is a statistical analysis of human decisions we must set our goals straight: statistics can help identify relations between variables and occurrences. it cannot help us determine individual personality biases in decisions like we are facing.
i.e. - it can help us see that loans for car has usually lower interest rates than loans to establish a small business - it is not in the scope of this work to determine whether this is a psychological bias, an economical analysis of future revenues from car ownership or just car-owning mentality, it also cannot ascribe quantity to such factors, or other non-continuos variables are determining the outcome of the decision

To tackle this problem there has to be some steps made:

1. identify orthogonal and dependent variables
   This will not be counted for further analysis
2. identify the count and noncount variables
   Factors will try to be explained and their effect demonstrated. They will not be used for the quantitative evaluation of the data.
3. Run a regression for the selected variables
   to determine quantitative effect.

Those are the variables given to try and explain the difference in Interest Rates approved

1. Amount.Requested
2. Amount.Funded.By.Investors
3. Loan.Length
4. Loan.Purpose
5. Debt.to.Income.Ratio
6. State
7. Home.Ownsership
8. Monthly.Income
9. FICO.range
10. Open.CREDIT.Lines
11. Revolving.CREDIT.Balance
12. Inquiries.in.the.Last.6.Months

13. Employment.Length

Orthogonal and dependent variables

some variable are interrelated and may affect the interst rate twice:\
the FICO score includes revolving debt data and so we should only use one of these variables (as they are heavily dependent)

variables that might be interrelated to the FICO score are 5,7,8,10,11,12,13
of those - 10,11,12,13 are specifically stated as included in the FICO score and when plotted against the Interest rate shows the same correlation line - I will ignore those variables for further calculation and use only the given FICO Range.

Personality Bias
Does a loaner from MN should get a lower interest over a loaner from AK? since this data IA, NC, PA, WI all get fairly low interests and DC and CO get relatively higher ones when examining same FICO groups - but this is not consistent in many same FICO groups and has no statistical meaning - this can be biased by so many human, social and local biases that can not be measured and the statistics are not the right measurement for.

In the same respective - does renting a house is better than owning a house with a mortgage? the Interest rate for people with different house ownership status shows no difference - this is also heavily biased by human factors and could not be effectively measured and calculated.

Loan purpose shows the same characteristics, and can not be tested thoroughly. There is a slight tendency by the board to give lower interest to people dealing with renewable energy, car, house, educational reasons, and higher interest to credit_card, debt_consolidation, other, small_business, wedding purposes. The human bias on this variable does not allow further inspection.

Visible Insights
--- put here some of the results shown in preliminery analysis

Data Massaging
In order to produce matrices valid for further analysis  only the accountable variables would be taken, and put into numeric formats. One variable would be added: since the amount of the loan approved does not matter for itself (you know if it is a relatively high or low amount by the requested amount) - the interest goes to the places where the loaners club did not approve the full amount requested by the loaner.
A new variable is added named ‘amountDiff’ that will show the percentage of the difference between the requested amount and the approved amount, distributed by the requested amount.

**Correlation and regression**

The new table made for analysis with the afforehead variables has been give the name: analysisTable

on this table one can calculate correlations between variables and run regression model. for the sake of this analysis I will not run further check-ups for moderation, mediation and significance but will only try to describe the data given.

A correlation Matrix will allow us to examine the relations between any 2 particular variables.
The matrix here shows on the Interest Rate given by the club

1. Very strong negative relation to the FICO score (0.7, 1 is a perfect correlation) The higher the FICO score is - the lower the interest is
2. A strong positive correlation (0.44) to the Loan length (longer time - higher interest)
3. A strong positive correlation (0.33) to the Amount requested

Regression model shown effective for those 3 Variables.
statistic significance proven and

Correlation inside same FICO groups - shows very high correlation (0.7) for both loan length and loan amount requested, far behind are monthly income and employment length at ~ 0.2 correlation.

**Test - re-test**

when running the same test over two random samples of half of the population - the correlation and significance remained the same.

**Used methods**

1. Correlation and Dependence - Wikipedia
2. Regression - Wikipedia
3. Test- retest (repeatability) - Wikipedia