Introduction:

A study was carried out regarding peer-to-peer loans issued through the Lending Club[1]. Lending Club is an online community bringing borrowers and lenders together with the stated goal of benefiting both by cutting the middle man broker (bank or finance) out and thus improving the terms of the loans for either party. As a result the borrowers pay less than if they had taken a loan from the bank and the lenders receive more than if they had deposited in a bank account. The interest rate of these loans is determined by the Lending Club on the basis of characteristics of the person asking for the loan such as their employment history, credit history, and creditworthiness scores.

The purpose of the analysis is to identify and quantify associations between the interest rate of the loan and the other variables in the data set. We expect to find a strong relationship between the interest rate and the credit worthiness (as measured by FICO score) of the borrower[2]. But we also consider whether any of the other variables have an important association with interest rate after taking into account the applicant's FICO score. For example, if two people have the same FICO score, can the other variables explain a difference in interest rate between them?

Methods:

Data Collection

We use data downloaded on **2500 sample loans [3]** (Each unique loan is in a separate row) made through the Lending Club. Each loan has observations for **14 different variables** each of which is in a separate column. The Variables have descriptive names (for example: `Interest.Rate` or `Loan.Length`) but the complete details for each variable can be found here [4]. We download the data and carry out the data analysis using the R programming language [5]

Exploratory Analysis

Exploratory analysis was performed by examining tables of the observed data. We identified transformations to perform on the raw data 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 other variables. Of the 35000 observations that should be present in the dataset there are **7 that are missing (i.e have NA value).** These occur in rows 367 and 1595 (not counting header as a row). If we remove these records (2 out of 2500) that have most of their numeric observations NA should not affect statistical modeling and have been removed from further data analysis. **There are also 77 occurrences of "n/a"** in the `Employment.Length` field. Assuming "n/a" means that the person is currently unemployed, these then have been changed to "< 1 year" for further factor analysis.

Statistical Modeling

The raw data contains FICO scores for the applicants that are in intervals of 5 (e.g: 785-789). This results in an unnecessarily large number of factors detracting from meaningful regression analysis. In order to do meaningful factor based analysis **FICO scores are converted to 4 factors in the following ranges (642,689] (689,737] (737,785] (785,832]** and used for the remainder of the study. This enables us to conduct regression analysis for a subset of data for each FICO range and identify the variables that show strong correlation with `Interest.Rate`. Finally after identifying the variables most strongly correlated to `Interest.Rate`, we will run multivariate regression [6] to describe `Interest.Rate` in terms of these variables.

Reproducibility

The data transformations needed on raw data are explained in the above sections. Further we read the loansData.csv (with `header=TRUE` and `as.is=TRUE`) file instead of the loansData.rda file which enabled us maximum flexibility in creating new factors from character fields (especially for FICO scores). Also we removed the row names that are present in the raw data. Readers are encouraged to ask for the R code from the writer to ensure full reproducibility.

Results:

As is expected there is a strong correlation between the FICO scores and the interest rate changed on the loan. In particular running regression on the factor based model we get estimates for coefficients (that are statistically very significant)

Coefficients:

```
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  16.2302     0.1038  156.33   <2e-16 ***
as.factor(loansData$FICO_4_ranges)(689,737]  -3.9060     0.1396  -27.97   <2e-16 ***
as.factor(loansData$FICO_4_ranges)(737,785]  -7.4970     0.1947  -38.50   <2e-16 ***
as.factor(loansData$FICO_4_ranges)(785,832]  -8.3320     0.3450  -24.15   <2e-16 ***
```

The formula for Interest rate can then be written as:

Interest.Rate(%) = 16.23 - 3.91¶`(FICO Range = "(689,737]")` - 7.5¶`(FICO Range = "(737,785]")` -8.33¶`(FICO Range = "(785,832]") + ei`

Ignoring the residuals, the range of values for `Interest.Rate` in the above equation is (7.9% to 16.23%). The range of values for the `Interest.Rate` from our dataset is (5.42% to 24.89%). Clearly the above equation does not explain all the variation in the Interest rate due to the FICO scores and there are some other factors that may lead to higher or lower Interest rates after controlling for the FICO score.

In order to get insight into `Interest.Rate` controlling for FICO scores, four independent data.frames (R data sets) for each of the 4 FICO score ranges are created. Regression analysis is carried out between `Interest.Rate` and each of the other variables independently for each of the FICO score ranges. The results of the regression analysis is shown below. If statistically significant correlation was found, the coefficient value and significance level is noted in the table below:

Table 1: Regression Coefficients for Interest.Rates against other variables in each of the 4 FICO Ranges

	Amount Requested	Monthly Income	Revolving Credit Balance	Loan Purpose	Employment Length	# of Inquiries in last 6 mos.	Debt to Income	# of Open Credit Lines	State (ref=CA)	Loan Length***
First quartile (642-689)	No Correlation	No Correlation	No Correlation	No Correlation	No Correlation*	0.150 (S @ 5% level)	0.038 (S @ 1% level)	0.061 (S @ 1% level)	VT,NH,DC,AK paid more	4.633 (S @ .1% level)
Second quartile (689-737)	No Correlation	No Correlation	No Correlation	Risky loans charged 2% more**	No Correlation*	0.4745 (S @ .1% level)	0.028 (S @ 5% level)	No Correlation	NM,AK paid more	4.83 (S @ .1% level)
Third quartile (737-785)	No Correlation	No Correlation	No Correlation	No Correlation	No Correlation	0.4817 (S @ .1% level)	No Correlation	No Correlation	Ks,MI,UT paid more	3.44 (S @ .1% level)
Fourth quartile (785-832)	No Correlation	No Correlation	No Correlation	No Correlation	No Correlation	0.4402 (S @ 5% level)	No Correlation	-0.156 (S @ .1% level)	WI,CO paid more	2.55 (S @ .1% level)

* Statistically Significant positive coefficient for employment > 10 years
** Small business or debt consolidation loans were charged 2% more interest rates at statistically significant levels.
*** Besides FICO score, loan length was strongly correlated to Interest Rates. For example people in lowest quartile paid on Average 4.633% higher interest rates for 60 months as compared to a 36 month loan.

As is immediately evident from the table, **besides FICO score only two other variables (`Loan.Length` and `Inquiries.in.last.6.Months`) consistently show strong correlation** with `Interest.Rate`. This becomes even more pronounced in the top two quartiles of FICO score ranges where there is no strong correlation to any other variable. This relationship can be succinctly seen graphically in the Figure1 attached to this writeup. For applicants in the bottom two quartiles, their current debt to income ratio and # of open credit lines is slightly correlated to the interest rates charged. Interestingly applicants in the bottom quartile that had been employed for more than 10 years tended to pay more Interest rates than other applicants. Also the purpose of the loan seemed to matter more for the applicants in the bottom two quartiles, with more risky loans for small business or debt consolidation charged higher interest. Using "CA" as reference we can see than people is some states paid higher interest rates. These however changed with FICO score ranges and hence no statistical inferences can yet be drawn and we are unable to predict a strong bias against or for residents of any state.

Loans Length: `Loan.Length` was very strongly correlated to interest rates. If the loan length changed from 36 months to 60 months, the interest rates charged changed as well. However people in higher quartiles of FICO score ranges paid lower additional interest rates for a 60 month loan (as compared to a 36 month loan) than did people the lower quartiles of FICO score range.

# of Inquiries in the past 6 months: This was also a strongly correlated variable but at higher significance levels (5% in some cases). Interestingly the # of inquiries did not affect the interest rates for people in the lowest quartile for FICO score as much as it did for people in there other three quartiles. Intuitively the "high risk" borrowers were already paying higher interest rates that the # of inquiries did not increase the risk by a large degree. This perceived increased risk however penalized the "low risk (i.e high FICO score borrowers)" more.

Confounders: Confounders are variables that are correlated to both the outcome and the covariates. They are hard to detect but may affect analysis. Since Interest.Rate depends on FICO scores and FICO score themselves are qualitatively based on other variables (such as `Debt.To.Income` or `Inquiries.in.Last.6.Months` or `Revolving.CREDIT.Balance)` intuitively we expect confounders to exist in our analysis. However since `Interest.Rate` is mostly explained by FICO scores and `Loan.Length` we deem influence of confounders in our analysis to be small and do not try to extract them.

If we run a regression analysis between `Interest.Rate, Inquiries.in.last.6.Months, Loan.Length` and FICO ranges, we get the following:

```
    Call:
lm(formula = loansData$Interest.Rate ~ loansData$Inquiries.in.the.Last.6.Months + as.factor(loansData$Loan.Length) + as.factor(loansDat$FICO_4_ranges))

    Residuals:
           Min      1Q  Median      3Q     Max
       -9.5066 -1.6822 -0.1513  1.6721  9.9427

    Coefficients:
                                                      Estimate Std. Error t value Pr(>|t|)
           (Intercept)                                15.03859    0.09723 154.666  < 2e-16 ***
           loansData$Inquiries.in.the.Last.6.Months    0.27968    0.04114   6.798 1.32e-11 ***
           as.factor(loansData$Loan.Length)60 months   4.44928    0.12162  36.584  < 2e-16 ***
           as.factor(loansData$FICO_4_ranges)(689,737] -3.98125    0.11246 -35.402  < 2e-16 ***
           as.factor(loansData$FICO_4_ranges)(737,785] -7.51080    0.15652 -47.988  < 2e-16 ***
           as.factor(loansData$FICO_4_ranges)(785,832] -8.38224    0.27655 -30.310  < 2e-16 ***
```

Thus according to this regression a person looking for a 60 months loans, in the lowest quartile of FICO scores and having 10 # of Inquiries in the last 6 months on average can expect to pay an interest rate of 15.04 + 4.45 + 0.28*10 = 15.04 + 4.45 + 2.8 = 22.29%. The difference between Interest Rate calculated by the regression equation and the actual observation is the residual. The standard deviation of residuals of the regression is 2.51 and the mean of residuals is approximately 0.

The 5% confidence intervals for the coefficients are given below. Since the t-statistics are large, as expected we see that the confidence intervals are tight around the coefficients:

```
                                                  2.5 %     97.5 %
(Intercept)                                 14.8479282 15.2292596
loansData$Inquiries.in.the.Last.6.Months     0.1990036  0.3603497
as.factor(loansData$Loan.Length)60 months    4.2107907  4.6877613
as.factor(loansData$FICO_4_ranges)(689,737] -4.2017720 -3.7607285
as.factor(loansData$FICO_4_ranges)(737,785] -7.8177174 -7.2038897
as.factor(loansData$FICO_4_ranges)(785,832] -8.9245299 -7.8399581
```

Finally we can write the equation for the `Interest.Rate(%)` as

Interest.Rate(%) = 15.03 +0.28(Inquiries.in.Last.6.Months) +4.45¶ *(Loan Length = "60")* - 3.98¶ *(FICO Range = "(689,737]")* - 7.5¶ *(FICO Range = "(737,785]")* -8.38¶ *(FICO Range = "(785,832]") + ei*

ei ~ (0,(2.51)^2)

Conclusions

Our analysis suggests that for the 2500 loans made by the Lending Club there is a statistically significant association between `Interest.Rate` and `Loan.Length`, FICO scores and `Inquiries.in.Last.6.Months.` Interest rates charged for the loans decrease as the FICO score increases. However since FICO score is a qualitative variable no linear relationship can be found. Instead we can relate `Interest.Rate` to ranges of FICO scores. Similarly the `Interest.Rate` depends on the `Loan.Length,` with borrowers paying more for the longer loans. Finally the `Inquiries.in.Last.6.Months` is linearly related to the `Interest.Rate`.

As part of the analysis it was noted that correlation of the variables change in different FICO score ranges, and indeed we may notice in some FICO score ranges weak correlation with other variables that are not part of the above equation show up. To reduce the residual error further and get a tighter fit for the equation for `Interest.Rate` may be separately written for each FICO range. This will enable variables that exhibit correlation in only some FICO ranges to be represented.

References

1. Lending Club. URL: https://www.lendingclub.com/home.action .Accessed 2/10/2013
2. Wikipedia "Credit Score in the United States" Page. URL: http://en.wikipedia.org/wiki/FICO_score#FICO_score .Accessed 2/10/2013

3. Lending Club 2500 Sample Loans Data. URL: https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv .Accessed 2/10/2013
4. Lending Club Loan Code Book. URL: https://spark-public.s3.amazonaws.com/dataanalysis/loansCodebook.pdf .Accessed 2/10/2013
5. R Core Team (2012). "R: A language and environment for statistical computing." URL: http://www.r-project.org/ .Accessed 2/10/2013
6. Seber, George AF, and Alan J. Lee. Linear regression analysis. Vol. 936. Wiley, 2012.