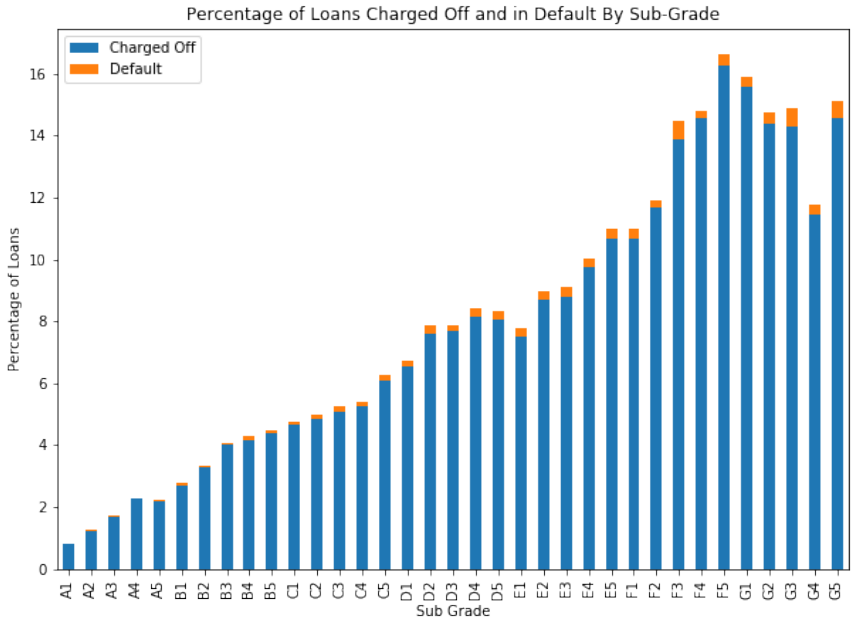
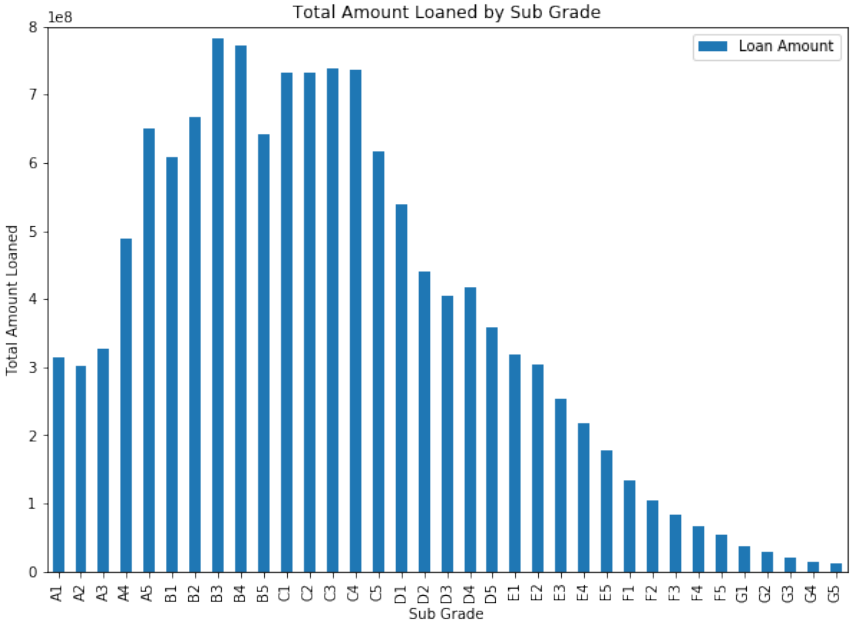
Application of Inferential Statistics to LendingClub Data
David Ermer

The LendingClub (LC) data has information about 887379 loans with 74 columns of data. The columns with FICO scores are missing and might contain the primary information used in the loan approval process. The LC process involved the ranking of the loans into 7 grades and 35 sub-grades. Calculating the percentage of charged off/in default loans in each of the 35 sub-grades shows that what ever process used by LC is reasonable and predictive of the performance of the loan.



LC also used their grading system to proportion the total amount loaned to each sub-grade.



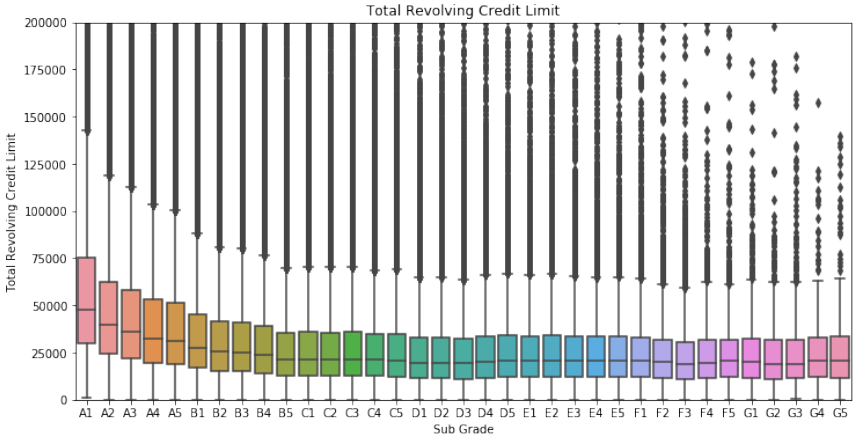
The provided data were processed and 38 columns were selected as being the most likely to contain information needed. As can be seen below there are a number of columns with 21372 rows of non-NaN data. This seems like it might be a refinement of other data columns, e.g. *Num Open Acc* (887350 entries) and *Num Open Acc 6mths* (21372 entries) although this can't be determined from the provided information.

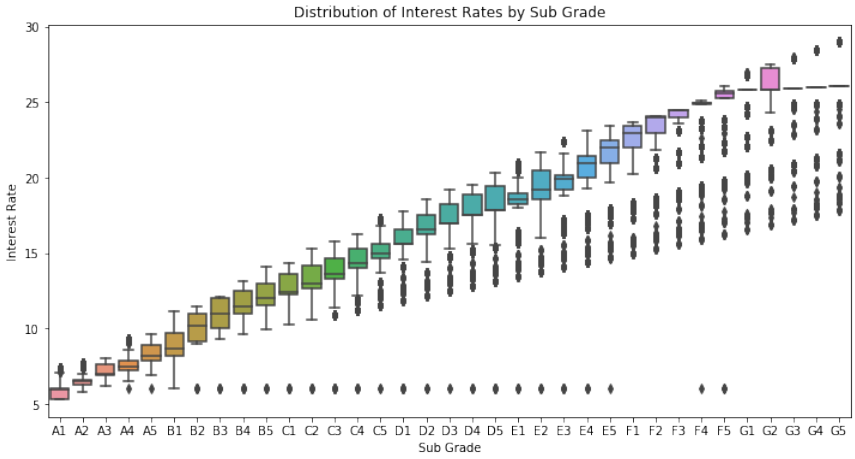
Sub Grade	887379
Loan Amount	887379
Interest Rage	887379
Annual Income	887375
Annual Income - Joint	511
Income Verification	887379
Income Verification - Joint	887379
Data Issued	887379
Debt Income Ratio	887379
Debt Income Ratio - Joint	509
Revolving Credit Balance	887379
Revolving Credit Ratio	886877
All Credit Ratio	21372
Current Deliquent Accounts	887379
Collections (12mths)	887379
Delquient (2yrs)	887379
Years Employed	887379
Home Ownership	887379
Inquiries (6mths)	887379
Personal Finance Inq	887379
Credt Inq 12mts	17389
Inst Credit Ratio	887379
Earliest Credit	887379
Max Bal Revol	21372
Total Coll E Owed	817103
Mths Last Delq	433067
Mths Last Derog	221703
Mths Last Rec	137053
Mths Last Inst	20810
Num Open Acc	887350
Num Open Acc 6mths	21372
Inst Open 12mths	21372
Inst Open 24mths	21372
Inst Open 6mths	21372
Open Revolve 12m	21372
Open Revolve 24m	21372
Total Revolve Limit	817103
Loan Status	887379

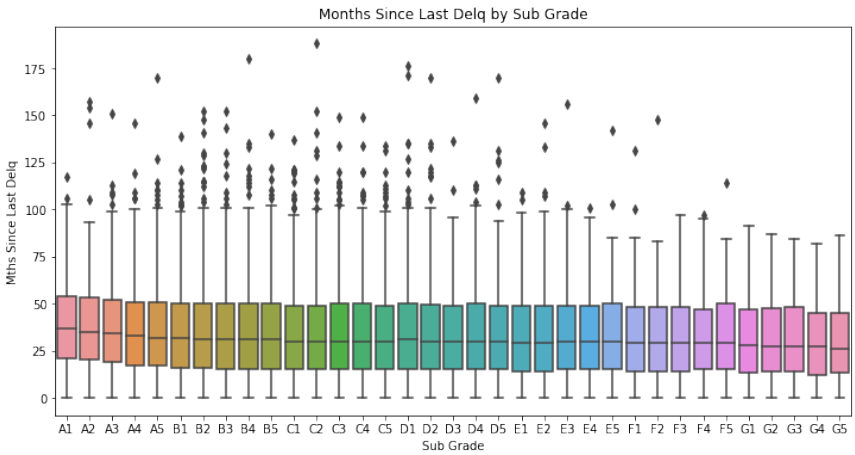
The data that is provided suffers from selection bias, i.e. it only includes data from loans that were approved. It maybe that the selection process has the effect for our purposes of homogenizing the statistics of the factors. Since the final goal is to generate a model to predict the performance of the loan it needs to be kept in mind that we are doing a 'second' selection process using homogenized data,

which will be a refinement of the original grading process done by LC and not an ab initio process and the results need to be interpreted in this light.

One question that needs to be determined is ‘are the differences in the factors across the different sub-grades statistically significant?’ Box graphs of most of our 38 columns grouped by sub-grade were made. For example:

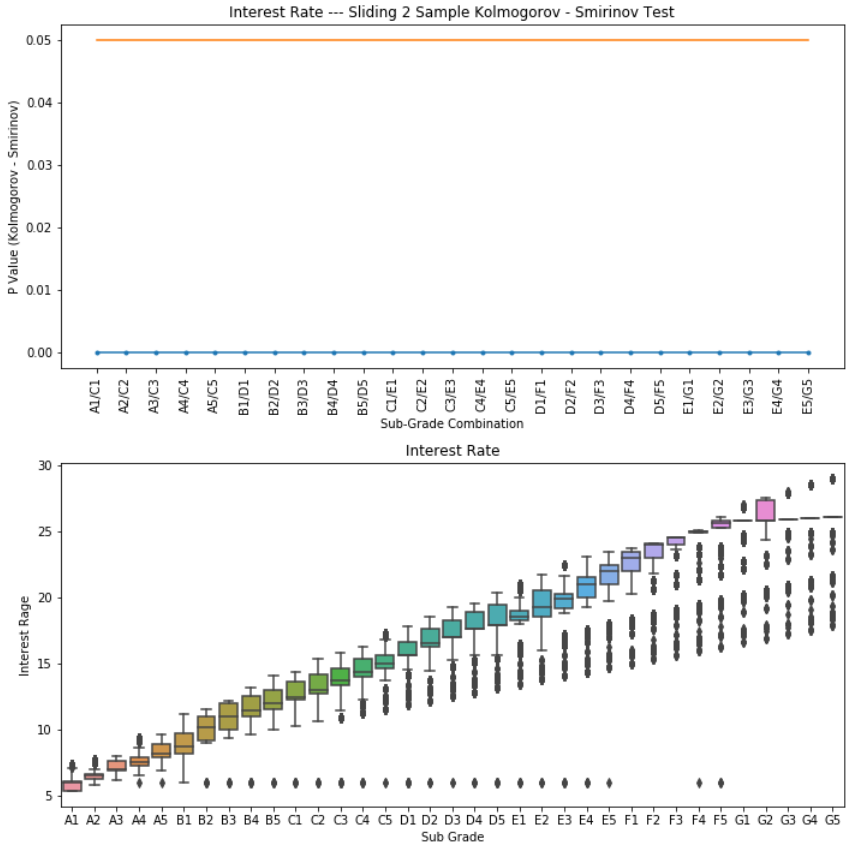






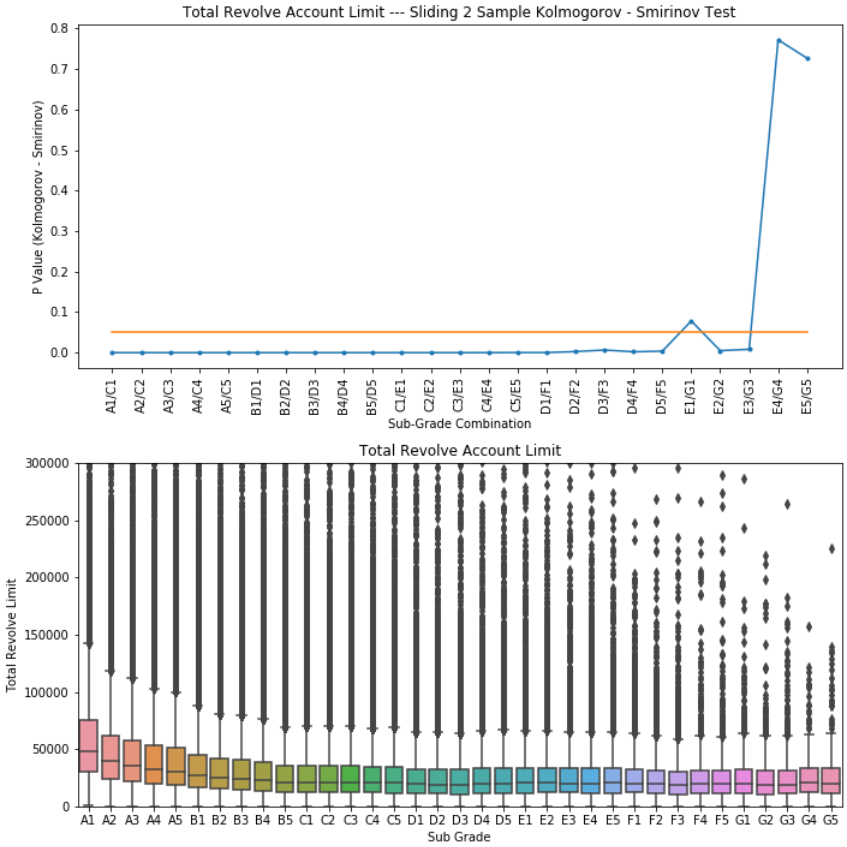
The interest rate is obviously assigned during and not a factor used in the loan approval process and it has as a result a strong correlation with the assigned sub-grade. The *Total Revolve Limit* and *Mths Last Delq* are both important factors in determining the sub-grade. However from the graphs it is not obvious that there is a significant difference in these factors among the different sub-grades. The mean of the credit limit changes by a factor of ~2, but the mean of the months since last delinquency does not vary much more than about 18%. However in both cases there is quite a bit of overlap between the distributions across all of the sub-grades. Most of the data in this data set have quite a few outliers.

The Kolmogorov-Smirnov test was selected to examine the statistical independence of the the factors across the sub-grades. The distribution of sub-grades spaced by 10 grades (i.e. two main grades) was tested and the p-value graphed for most of the columns. Other intervals were tried but similar results were obtained.

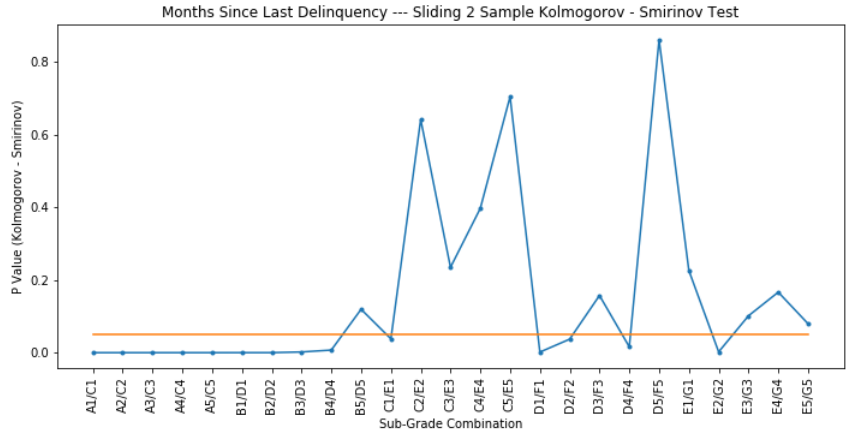


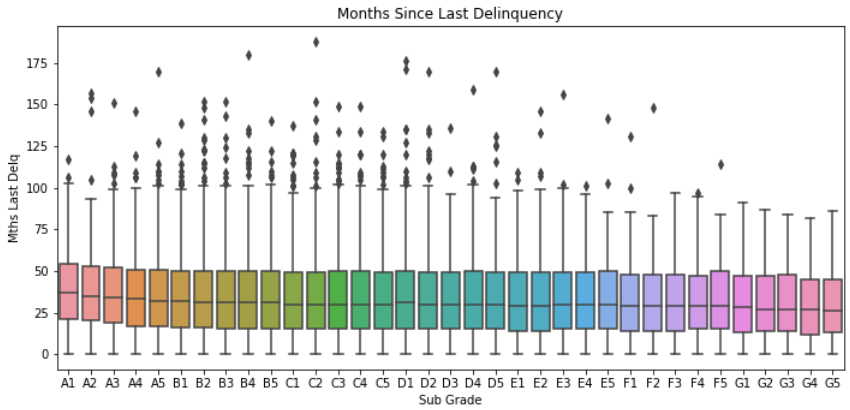
The interest rate obviously shows independence of the distributions across the sub-grades.

The distribution of the total revolving credit limit data is by the KS test significantly different across the sub-grades. There are many fewer instances at the lower sub-grades which may effect the test between those grades.



The *Mths Last Delq* data shows much less variance across the sub-grades and KS test results in a p-value above 0.05 across a range where there are actually a large number of instances in the sub-grade.





Most of the columns with numeric data were tested this way with various results. Some of the data are quite homogenized across the sub-grades while others are not.

This will be important information that can be used to select the most relevant data used in developing a model to predict the performance of loans based on the data provided in the data set.