**Tejsvi Rai – Case Study Write-up**

*Introduction:*

This case study was coded up in Python using the latest pandas, matplotlib, scikit-learn and numpy packages.

*Part 1: Data Exploration and Evaluation*

The first part of this assignment involved loading, plotting and analyzing the data. In my code, this is done by the following functions:

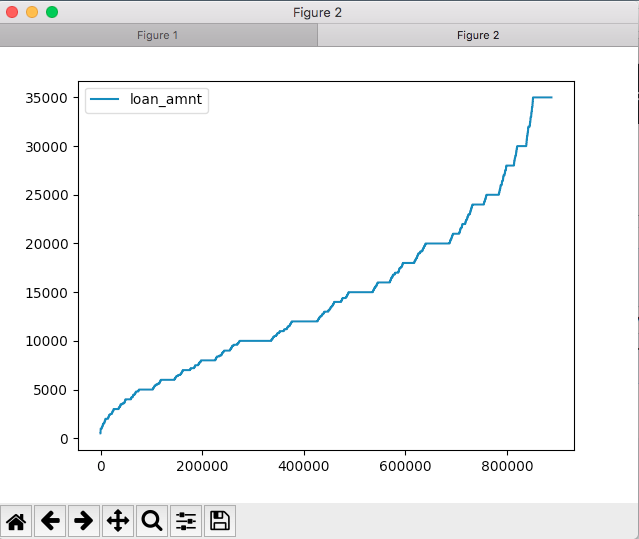
* Load\_raw\_data()
* augment\_data()
* visualize\_data()
* clean\_data()
* analyze\_data()

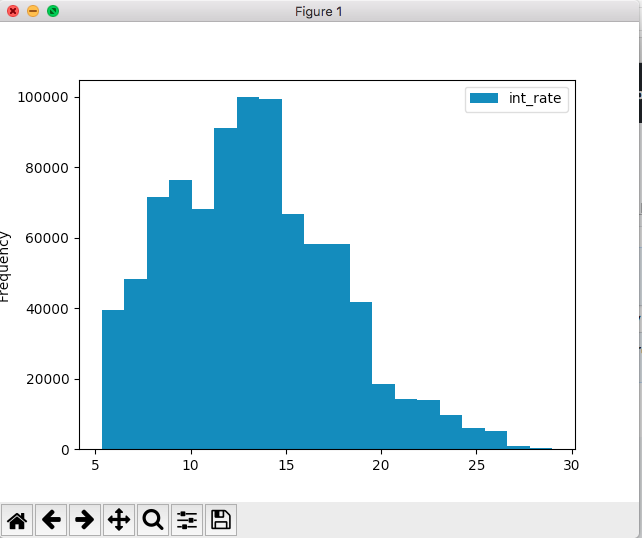
The load\_raw\_data() function simply loads the csv file containing the loan information into a pandas Dataframe. We have the option to load a subset of the data for faster testing

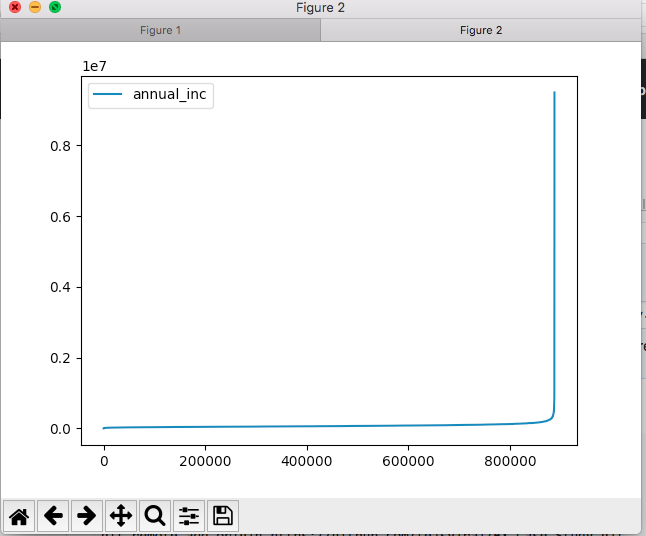
The augment\_data() function adds extra columns to the data that will later be useful for answering business related questions. It performs low level tasks such as converting loan grades into numeric values (needed for the logistic regression later on) or converting issue dates from strings to datetime objects (needed to filter out valid loans)

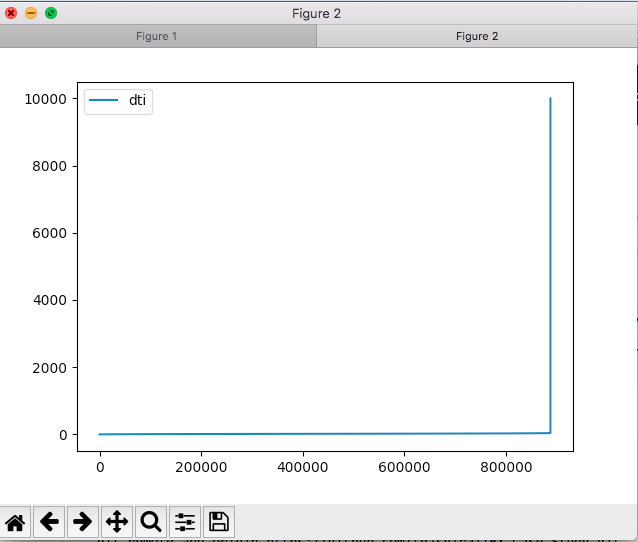
The visualize\_data() function either outputs subsets of the data onto the console or plots the data to detect outliers. From this function, we can see several things:

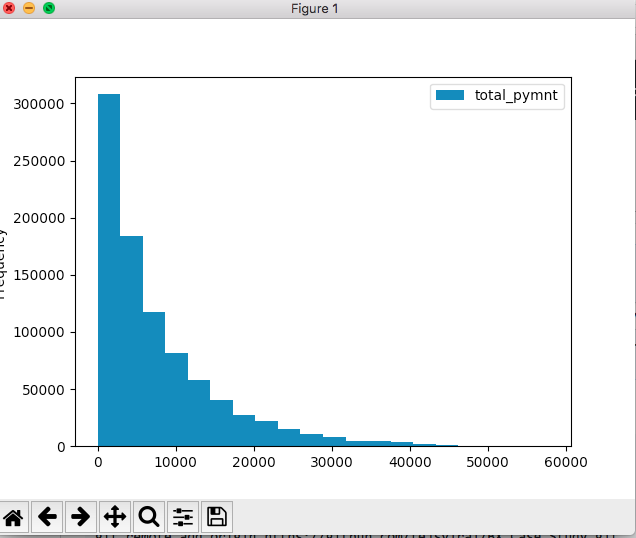
* The dataset is nearly complete; only one column (“annual\_inc”) has missing values (4 missing values out of 887379 rows)
* Most numeric columns have very clean data. Examples of some data integrity tests that a simple DataFrame.describe() function can answer is:
  + loan\_amnt > 0
  + funded\_amnt > 0
  + int\_rate > 0 and < something large like, say, 100%
  + annual\_inc >= 0
  + dti >= 0 and < something large like, say, 10x
  + revol\_bal >= 0
  + total\_pymt >= 0
  + funded\_amnt/ load\_amnt ratio >= 0 and <= 1
* We see from the summary statistics that the annual\_inc, dti and revol\_bal variables are highly skewed.
  + This is confirmed when we plot the data in either a sorted or histogram format.
* We also find a few clearly bad values in the dti data corresponding to zero annual income
* We find that the rest of the numeric columns are extremely well behaved in terms of their data distributions





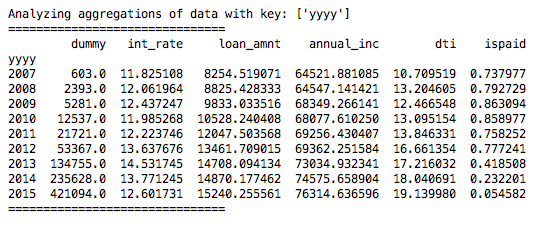


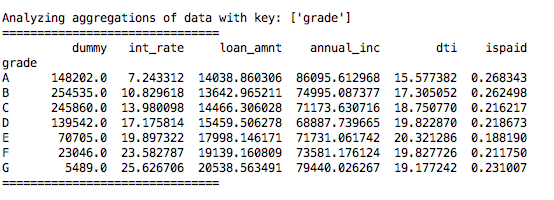


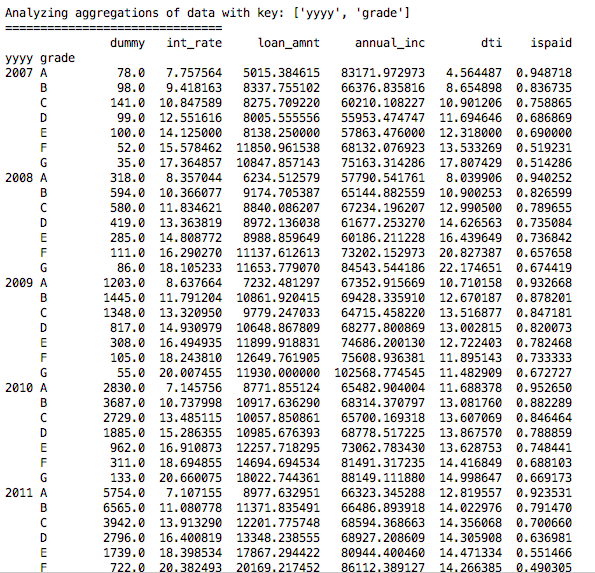


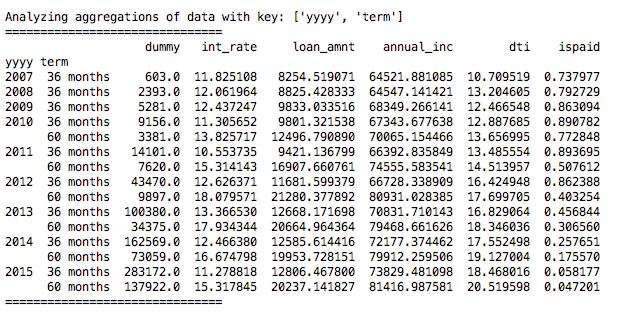
The clean\_data() function removes the bad dti values and winsorizes the three skewed columns at a very high percentile to provide a balance between preserving data integrity and allowing for more robust inference

The analyze\_data() function then produces summary statistics across a number of different data groupings. From this function, we can see that loan issuance has increased dramatically over time even as the interest rate on the loans has stayed relatively constant. Moreover, the firm tends to mostly make high quality loans (A – D grade), very few low quality loans (grades E – G) with a sweet spot in the B / C grades. Predictably, the average interest rate on the loans rises as the riskiness of the loan increases. Consistent with the above facts, the proportion of loans that eventually get repaid in each year by grade are highest for grade A and lowest for grade G. This indicates that the firm is appropriately pricing its loans for both perceived and actual riskiness. Finally, we notice that the firm originally only issued 36-month loans but has since expanded its business into the 60-month term as well.









Part 2: Business analysis questions:

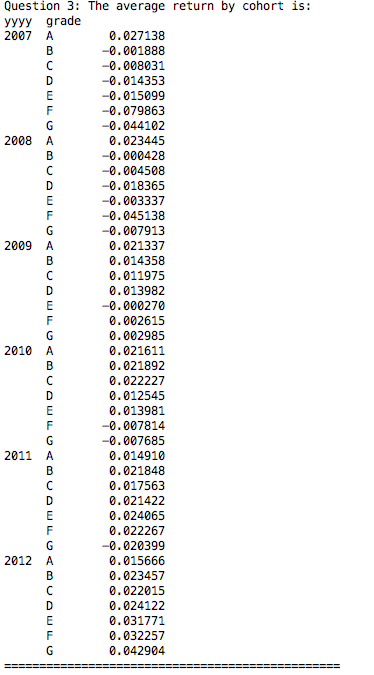
Thanks to the augment\_data() function, answering these questions was as simple as filtering the loans to only be completed 36-month loans and then doing a bunch of group\_bys in pandas and aggregating the results.

Business Analysis Output:

================================================

Question 1: The percentage of loans that has been fully paid is: 86.85%

Question 2: The most delinquent cohort is: (2007, 'G')



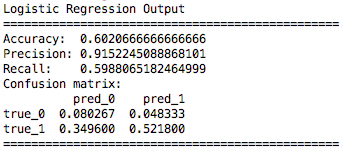
*Part 3: Logistic Regression*

This part relies heavily on the scikit-learn module to implement a simple logistic regression model. Here, the predictor variable is whether or not the loan was repaid while the independent variables are easily observable characteristics like loan quality (converted from letter grades to numeric values), interest rate, annual income and a combined dti + loan / income variable designed to capture a borrower’s total debt burden.

The features chosen are intuitive – loan quality, interest rates, a borrower’s ability to service the loan via her income and her existing debt burden are all germane to the credit quality of a particular loan and are, in fact, the primary tools used by credit officers in deciding whether to extend a loan.

The model is fitted using 80% of the data chosen randomly, with 20% held back for validation.

The metrics used for validation are standard in logistic regression literature – confusion matrix, accuracy, recall and precision:



The model has high precision, implying a pretty low false positive rate. This means that of the loans it classified as good loans, a high fraction of the loans were actually good. This is nice because it means that we can trust the model’s predictions when deciding if a loan is good or not.

On the other hand, the model has pretty average recall, implying a pretty high false negative rate. This means that of the universe of loans that are actually good, the model misclassified a fair amount as bad leading to missed investment opportunities.

To summarize, when the model says a loan is good, it is actually good with a high probability. However, a lot of the loans the model says are not good are actually good. This means that while we can reliably use the model for investment decisions, it doesn’t identify the entire universe of good loans as well as it could.