**Advances in Data Science/Architecture**

Assignment 2 – Part 1

Submitted By:

Karan Millan Thevar

Pratik Patre

Rahul Chandra

(Team 3)

17th March, 2017

Introduction

The Project focuses on 2 major sections i.e ‘Data wrangling and exploratory data analysis’ and ‘Building and evaluating models’. In this report only the first section is discussed. The focus is on Data retrieval, preprocessing, Feature Engineering, implementing pipeline, Docker and deriving a summary matrix out of it. This process consists of programmatically downloading data from Lending club data set using this url - <https://www.lendingclub.com/info/download-data.action> . Here, the list of all the loan data and declined data are to be downloaded, processed and analyzed. This analysis is then shown visually on PowerBI. The whole data is presented in a Luigi pipeline and then Dockerized. The Docker image is then hosted and fired from Amazon AMI EC2 instance. The output file will then be automatically posted on S3.

There are many Derived columns added to both the data set like:

1. Loan Data

FICO Score, UPB, Risk Score Percentage

1. Declined Loan data

Gross Income, IsEmployed

**Data wrangling**

Data wrangling is loosely defined as the process of converting or mapping data from one "raw" form into another format that allows for more convenient consumption of the data with the help of automated tools. This section is done using Python coding.

**1.A Deliverables for this section**

* Docker image for the whole process.
* Instruction to run docker image to replicate the whole process
* PowerBI dashboard which is shared through PowerBI public

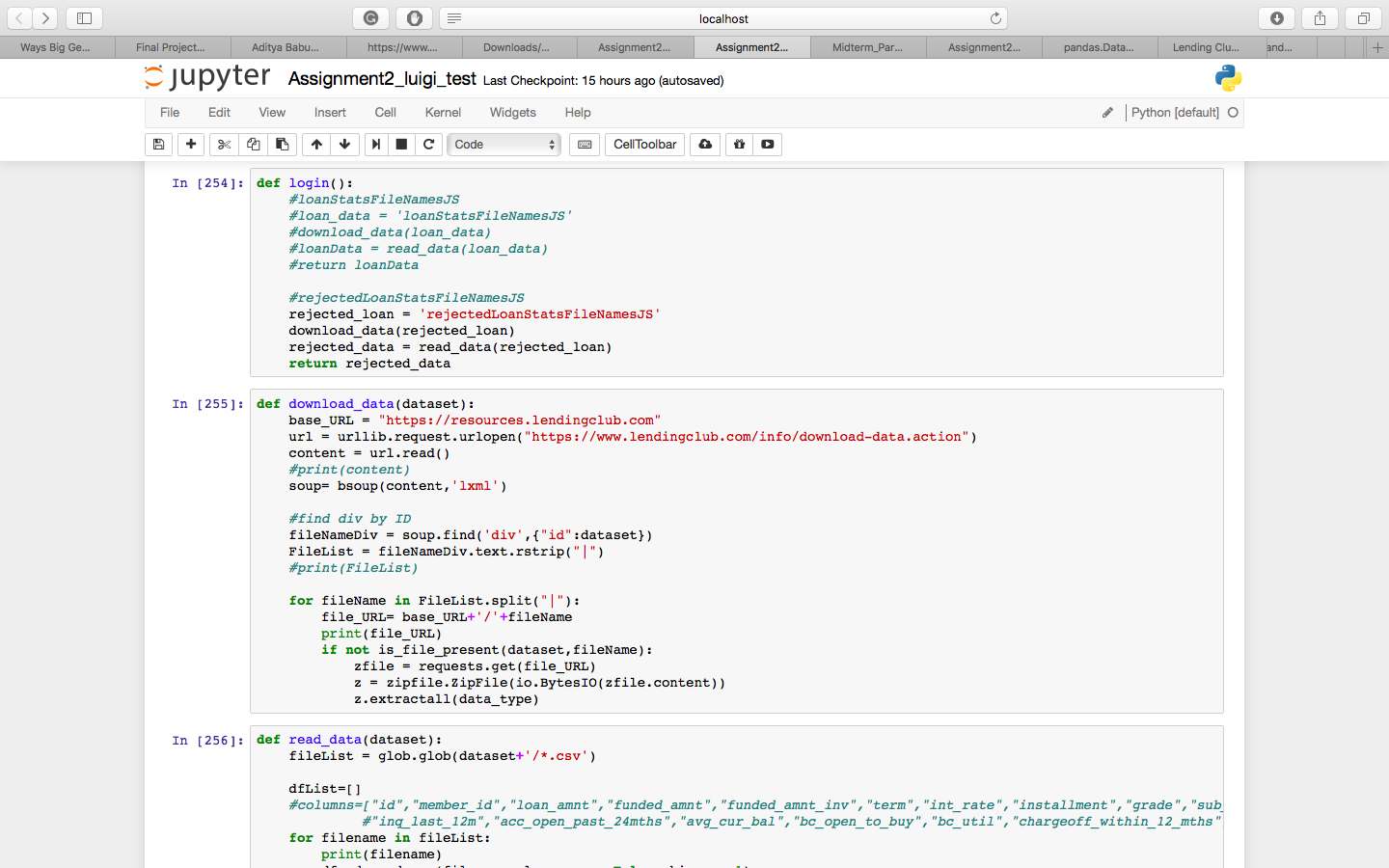
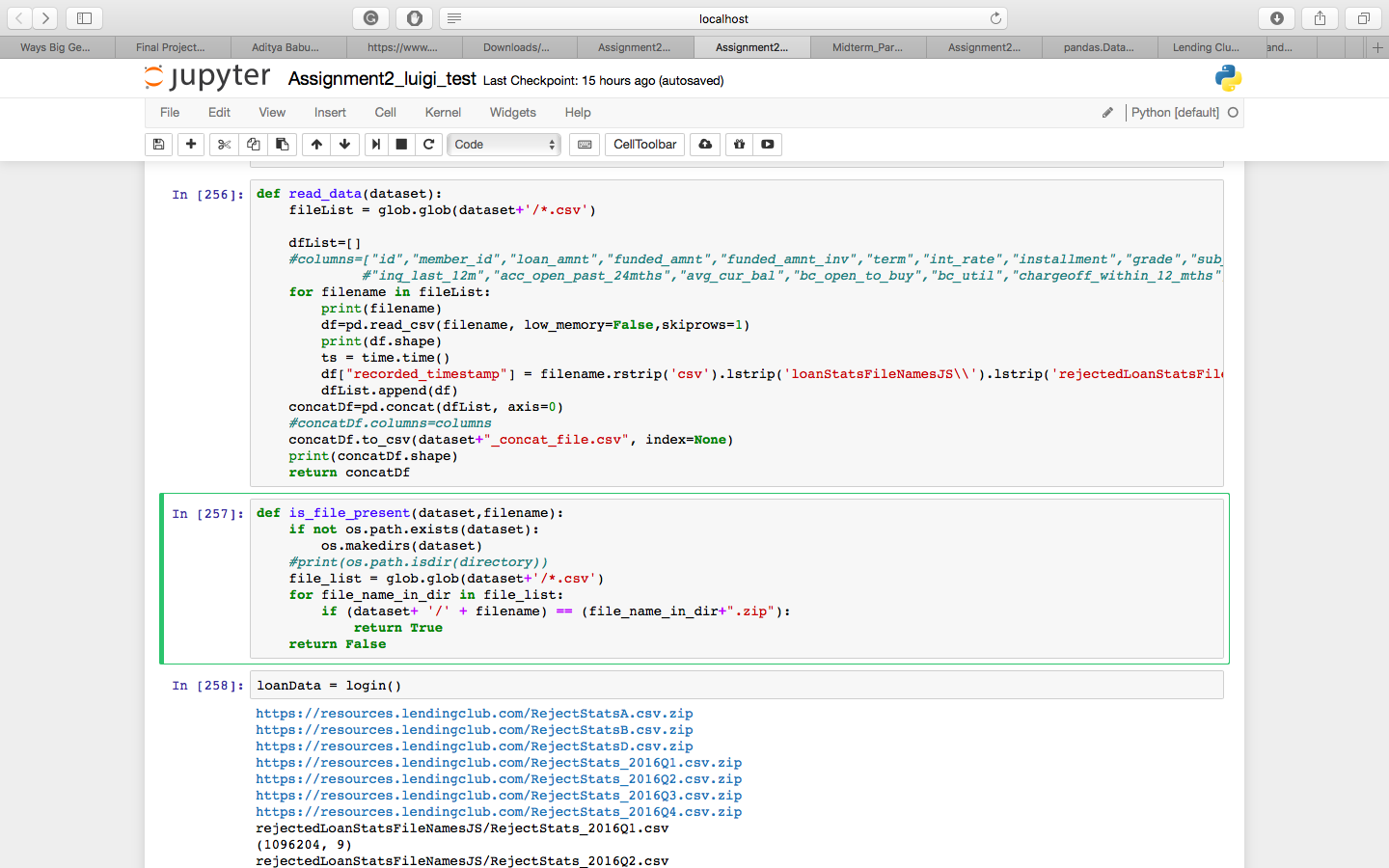
**1.B Process followed**

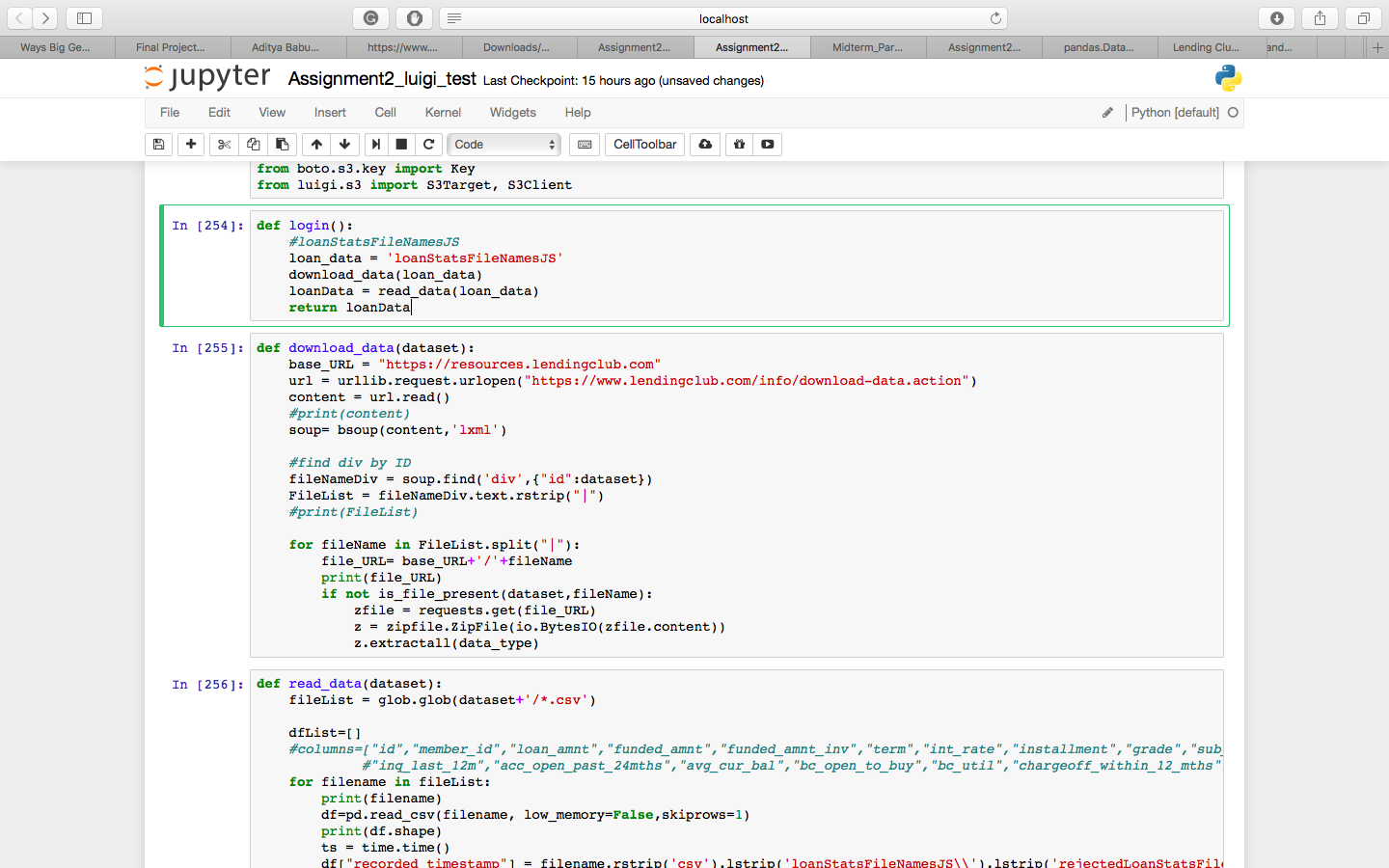
Steps followed to execute this section and generate the data programmatically without human inter-face.

**Step1: Scrapping Data from the web-site**

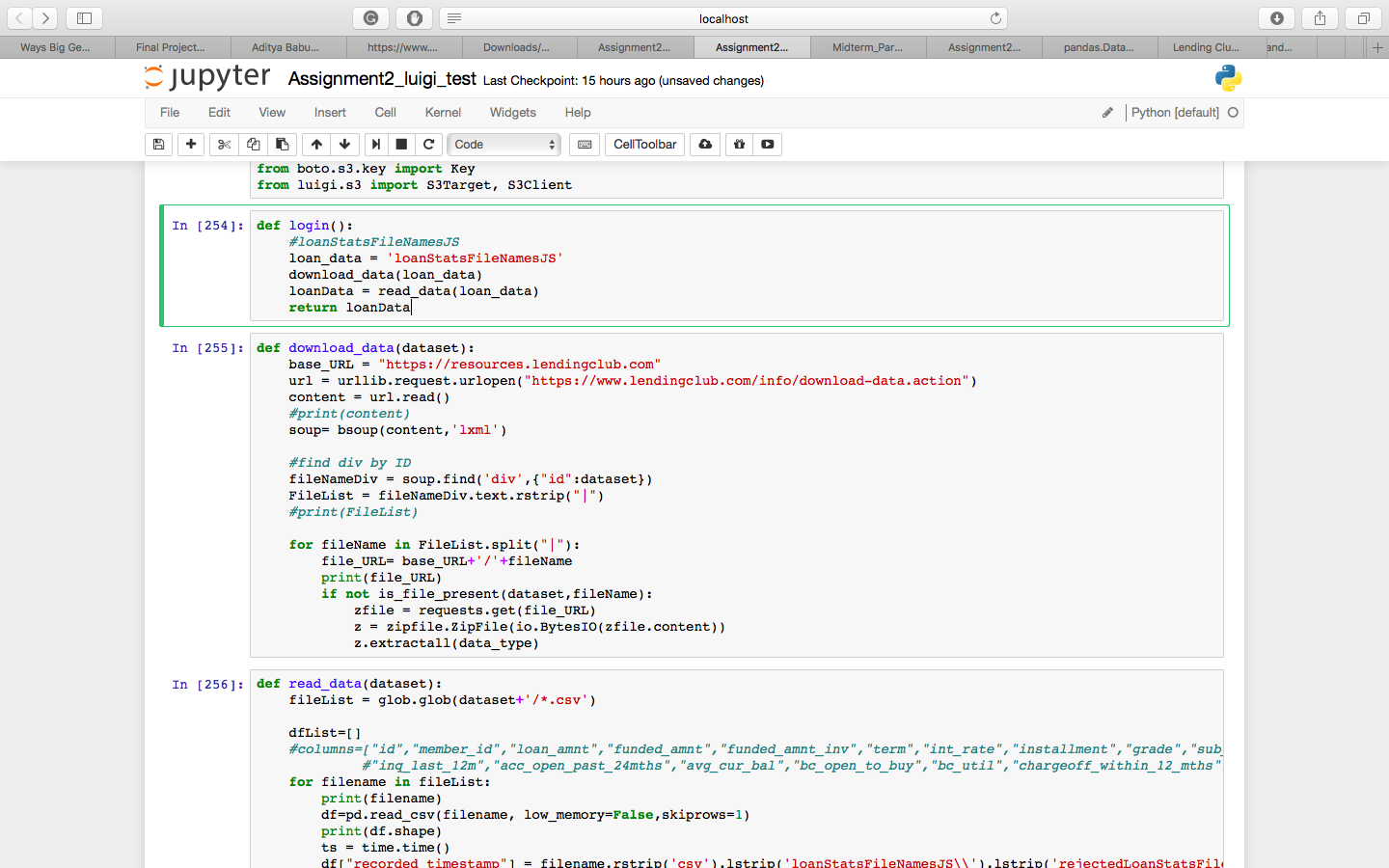
The first part of this section is to programmatically download two types of Loan Data from the url (<https://www.lendingclub.com/info/download-data.action>). We have used python to complete this section. Libraries like requests and BeautifulSoup are used to perform this process. Session data is saved and then the .csv files are extracted by accessing the div id and then getting the file names.

Below is the code snap shot that is used to perform this process:





The same process is followed for the Loan Data in the different Luigi pipeline



**Step 2: Scraping of data and Preprocessing it.**

After the scrapping of data is done the next major task is Preprocessing. The data in the csv downloaded file has many unwanted rows which do not provide any information. Thus, these rows are removed and then the data from different files are joined together.

This processed data is then saved for further use.

**Step 3: Missing Data handling**

The major columns which addressed during missing data handling were "Months since oldest bank installment account opened", "Months since oldest revolving account opened", "Months since most recent account opened", "Number of bankcard accounts", "The number of inquiries in past 6 months".

It was important to handle Months since oldest bank installment account opened as it is one of the key factors in deriving FICO Score. This column is handled by using values from Months since oldest revolving account opened as it is relevant by values. If the there is no value in oldest revolving account opened then the most logical other column would be Months since most recent account opened. If all the columns are NA then we are left with handling all the three values by inserting 1 month as value.

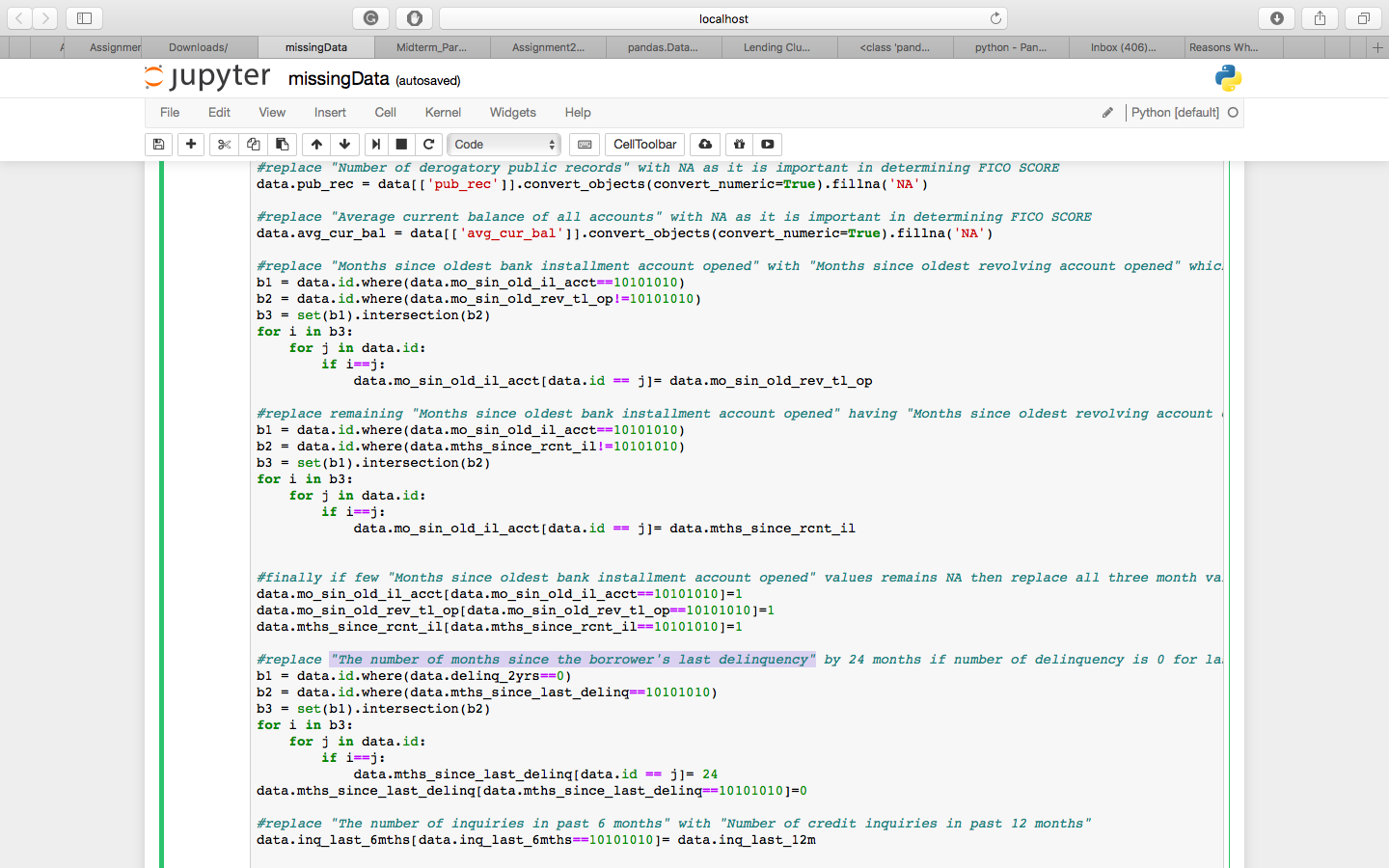
Number of bankcard accounts is another important factor for calculating FICO Score. This column is handled by using summation of number of revolving accounts and installment accounts. This information is provided by Official FICO Score website. It is total number of revolving and installment accounts.

"The number of months since the borrower's last delinquency" is handled as 24 for those accounts which have delinquency as 0 for last 2 years. The rest of the missing data is handled by inserting 0.

The number of inquiries in past 6 months is handled by number of inquiries in past 12 months.

"Number of derogatory public records" and "Average current balance of all accounts" is handled by replacing it with NA as it is important in determining FICO SCORE. The NA values come in a specific slab for which the points are allocated so it was important to have NA values rather than to replace it with any other value. This data is again fetched from FICO website.

Below is the code snippet for the above:



**Step 4: Feature Engineering**

The cleaned and processed data is now ready for Feature engineering. The primary task here is to get the required and appropriate features that would effect the interest rate. The decision to select the columns was done through generating the pearson correlation matrix. This gave us a wage idea to which column to take and which to reject.

Pearson Correlation was major factor in finding out dependency of various columns on Interest Rates. Few of the columns where coming as NaN which helped us in eliminating the columns.

Other columns like policy code, grades and sub grades are specific to the Lending Club company and so is irrelevant for data analysis here. It must be a derived attribute of the Lending Club company from their analysis. Since we are not aware of the analysis behind the attribute, it cannot be used in our analysis.

Keeping in mind the matrix there where some columns discarded from the whole list.

They are:

1. Grade

2. SubGrade

3. Emp Title

4. Desc

5. URL

6. Payment Plan

7. Zip Code

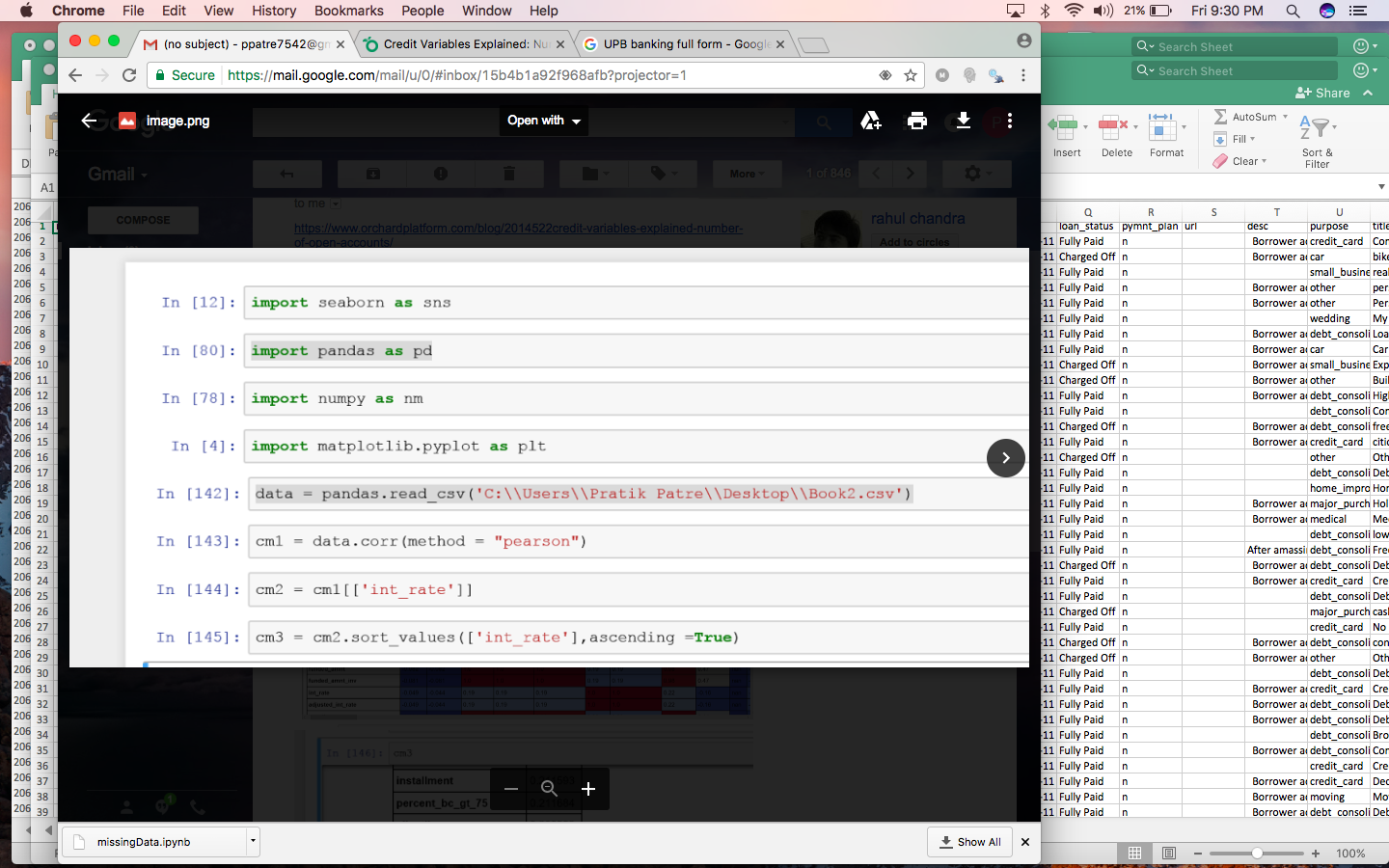
8. Title

9. Policy Code

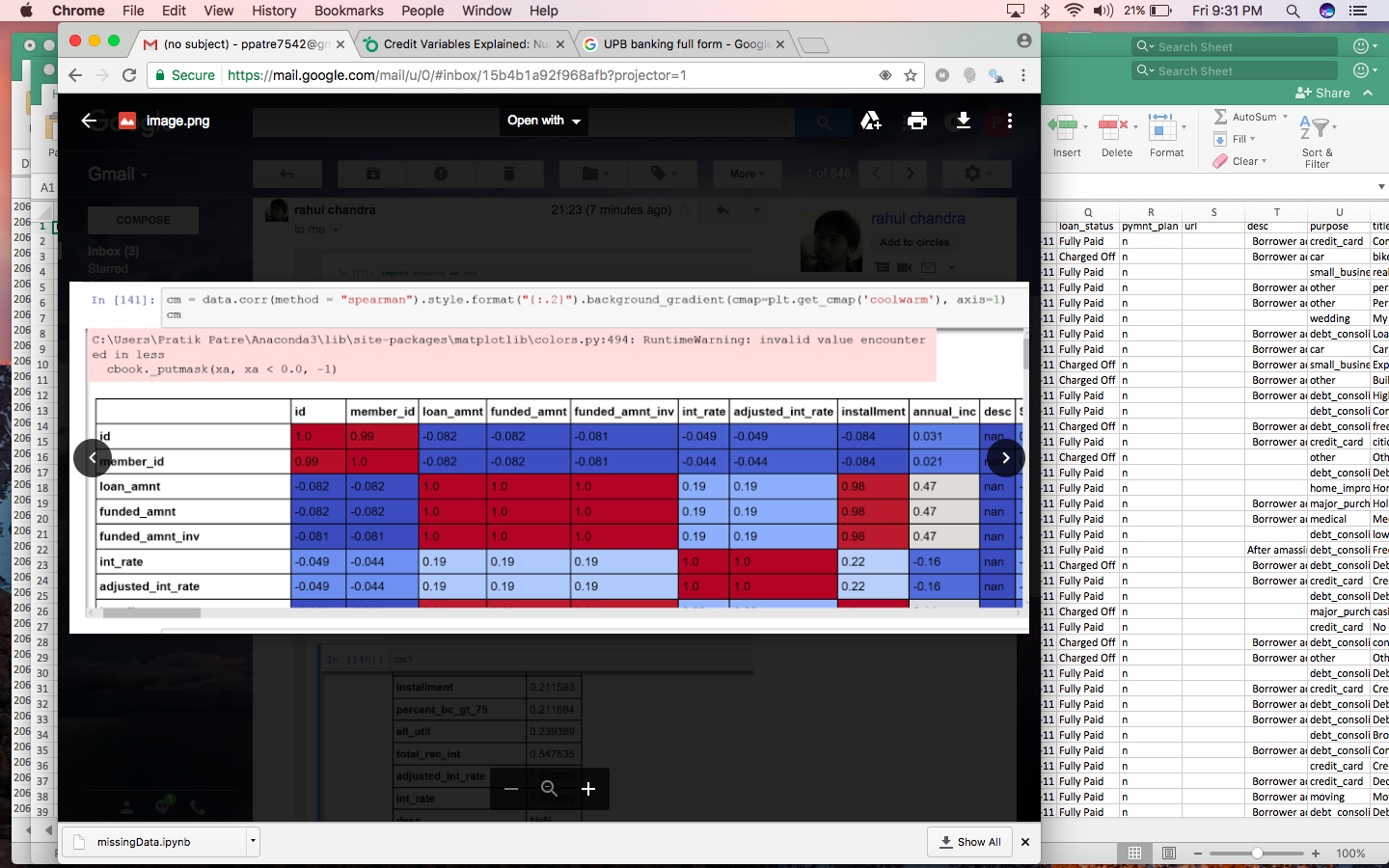
10. Verification Status Joint

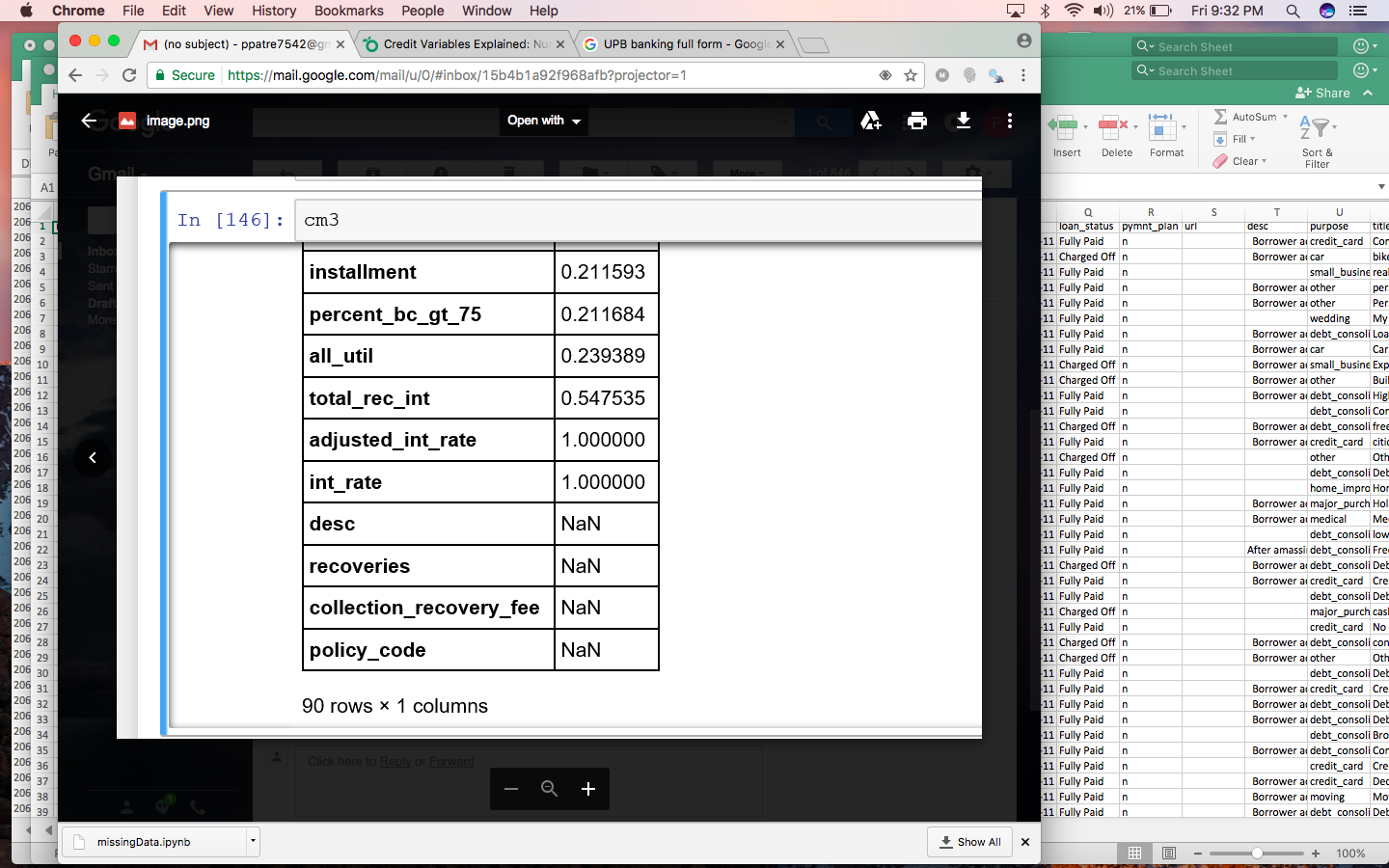
**Pearson correlation matrix**

Below is the code snippet:



Output matrix:





**Step 5: Derived attribute**

These are the new added column which give some sense to the data. They are derived from the combination of multiple attributes already present in the file.

After doing intensive analysis on the data and interest rate dependencies, we where able to calculate FICO Score, UPB, Risk Score Percentage.

**FICO SCORE**

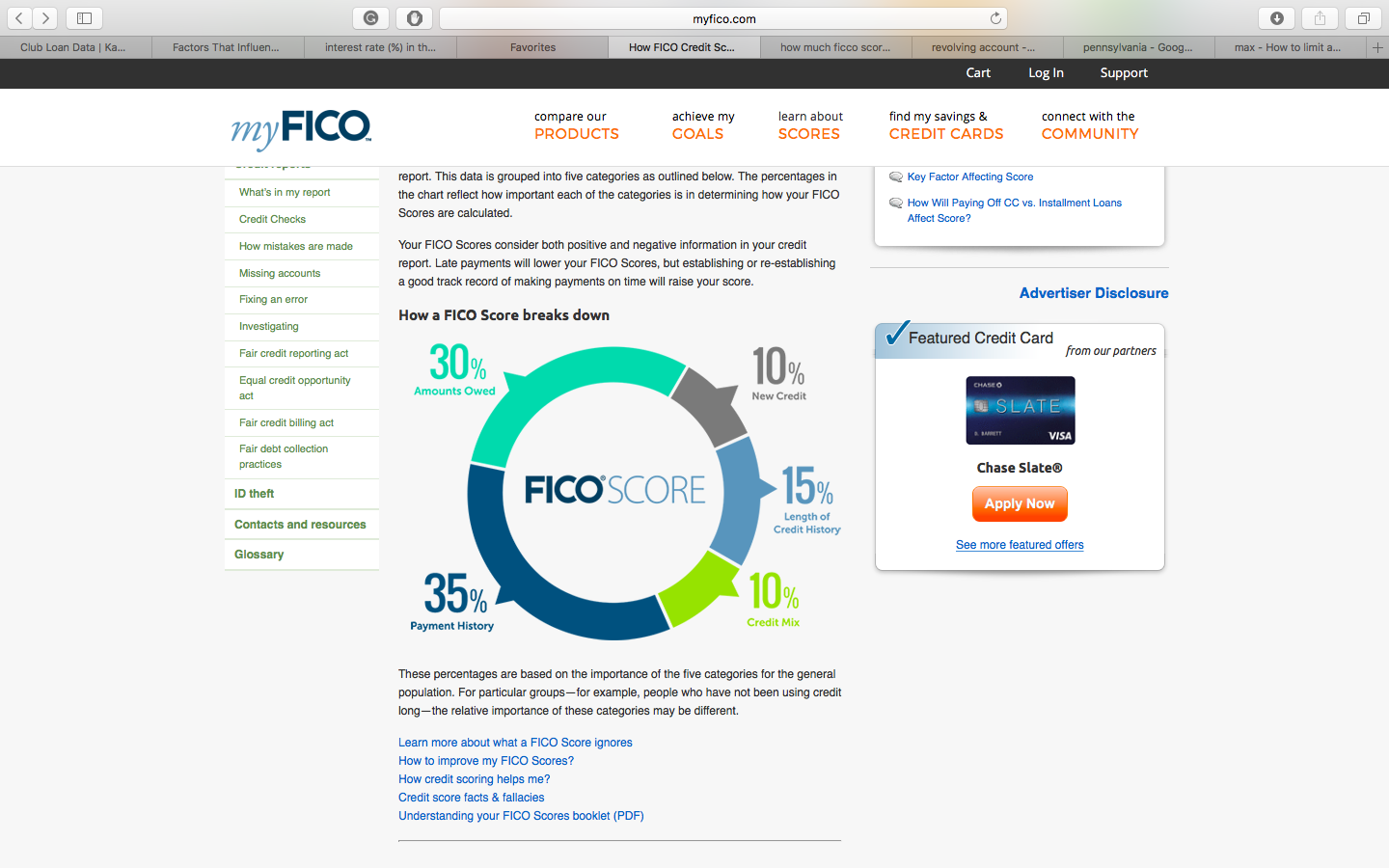
What is the 'FICO Score'?

A FICO score is a type of credit score created by the Fair Isaac Corporation. Lenders use borrowers' FICO scores along with other details on borrowers' credit reports to assess credit risk and determine whether to extend credit. FICO scores take into account various factors in five areas to determine credit worthiness: payment history, current level of indebtedness, types of credit used, length of credit history and new credit accounts.

FICO scores range between 300 and 850. In general, scores above 650 indicate a very good credit history. In contrast, individuals with scores below 620 often find it difficult to obtain financing at favorable rates. To learn more about how your credit score is calculated, read "Does Having Several Credit Cards Hurt My Credit Score?" To determine credit worthiness, lenders take a borrower's FICO score into account but also consider other details such as income, how long the borrower has been at his job and type of credit requested.

Calculating FICO Scores

To determine credit scores, the Fair Isaac Corporation weighs each category differently for each individual. However, in general, payment history is 35% of the score, accounts owed is 30%, length of credit history is 15%, new credit is 10% and credit mix is 10%.



Payment History

Payment history refers to whether an individual pays his credit accounts on time. Credit reports show the payments submitted for each line of credit, and the reports indicate if the payments were received 30, 60, 90, 120 or more days late.

Accounts Owed

Accounts owed refers to the amount of money an individual owes. Having a lot of debt does not necessarily equate to low credit scores. Rather, FICO considers the ratio of money owed to the amount of credit available. To illustrate, an individual who owes $10,000 but has all of his lines of credit fully extended and all of his credit cards maxed out may have a lower credit score than an individual who owes $100,000 but is not close to the limit on any of his accounts.

Length of Credit History

As a general rule of thumb, the longer an individual has had credit, the better his score. However, with favorable scores in the other categories, even someone with a short credit history can have a good score. FICO scores take into account how long the oldest account has been open, the age of the newest account and the overall average.

Credit Mix

Credit mix is the variety of accounts. To obtain high credit scores, individuals need a strong mix of retail accounts, credit cards, installment loans such as signature loans or vehicle loans, and mortgages.

New Credit

New credit refers to recently opened accounts. If the borrower has opened a bunch of new accounts in a short period of time, that indicates risk and lowers his score.

To obtain the FICO score we choose these variables from the column list:

1) Amounts Owned - avg\_cur\_bal num\_rev\_tl\_bal\_gt\_0 (number of rev trades with bal>0) total\_bal\_il revol\_bal recoveries (30%)

2) Payment History - pub\_rec (35%)

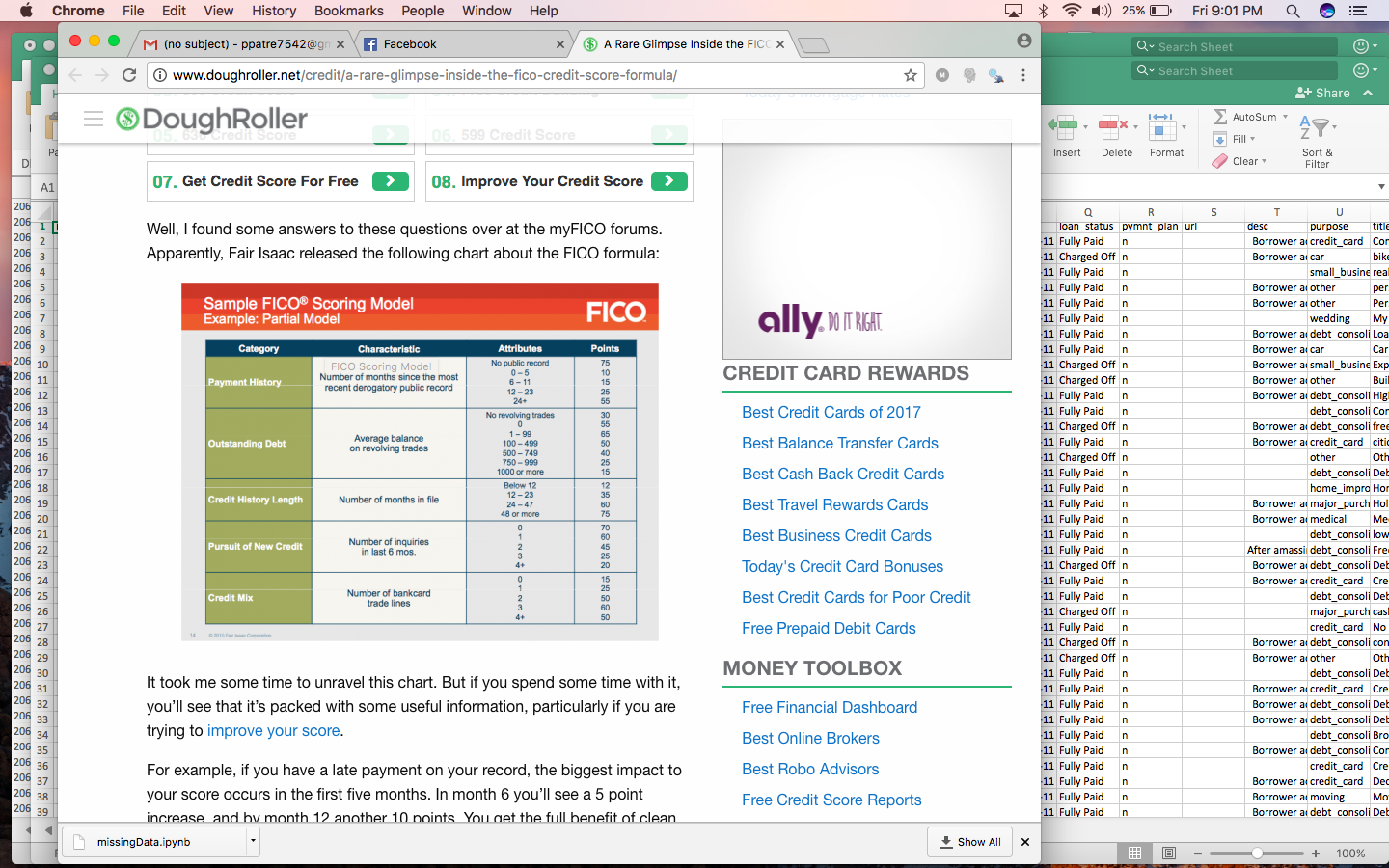
3) Length of Credit History - mo\_sin\_old\_il\_acct (15%)

4) Credit Mix - num\_bc\_tl (10%)

5) New Credit - inq\_last\_6mths (10%)

To use these values and then formulate them to get the exact FICO score is difficult. Thus, by creating our own formula inspired from the original, we creating slabs calibration for every attributes. This helped in easy calculation. The FICO score we got was in term of percent where 100% is the best score.

The calibration is given as follows:

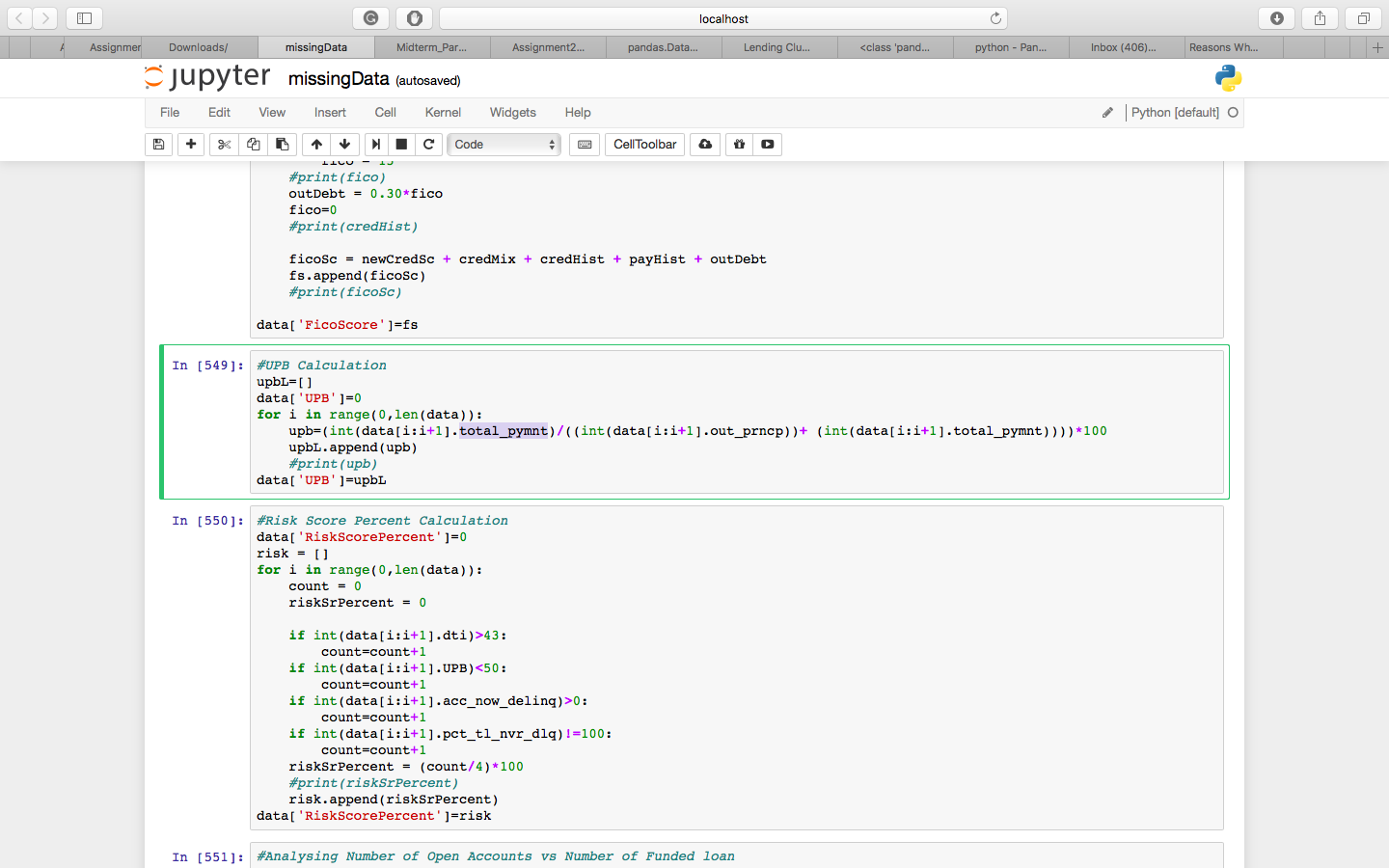


The code snippet for the same is given below:

**UPB**

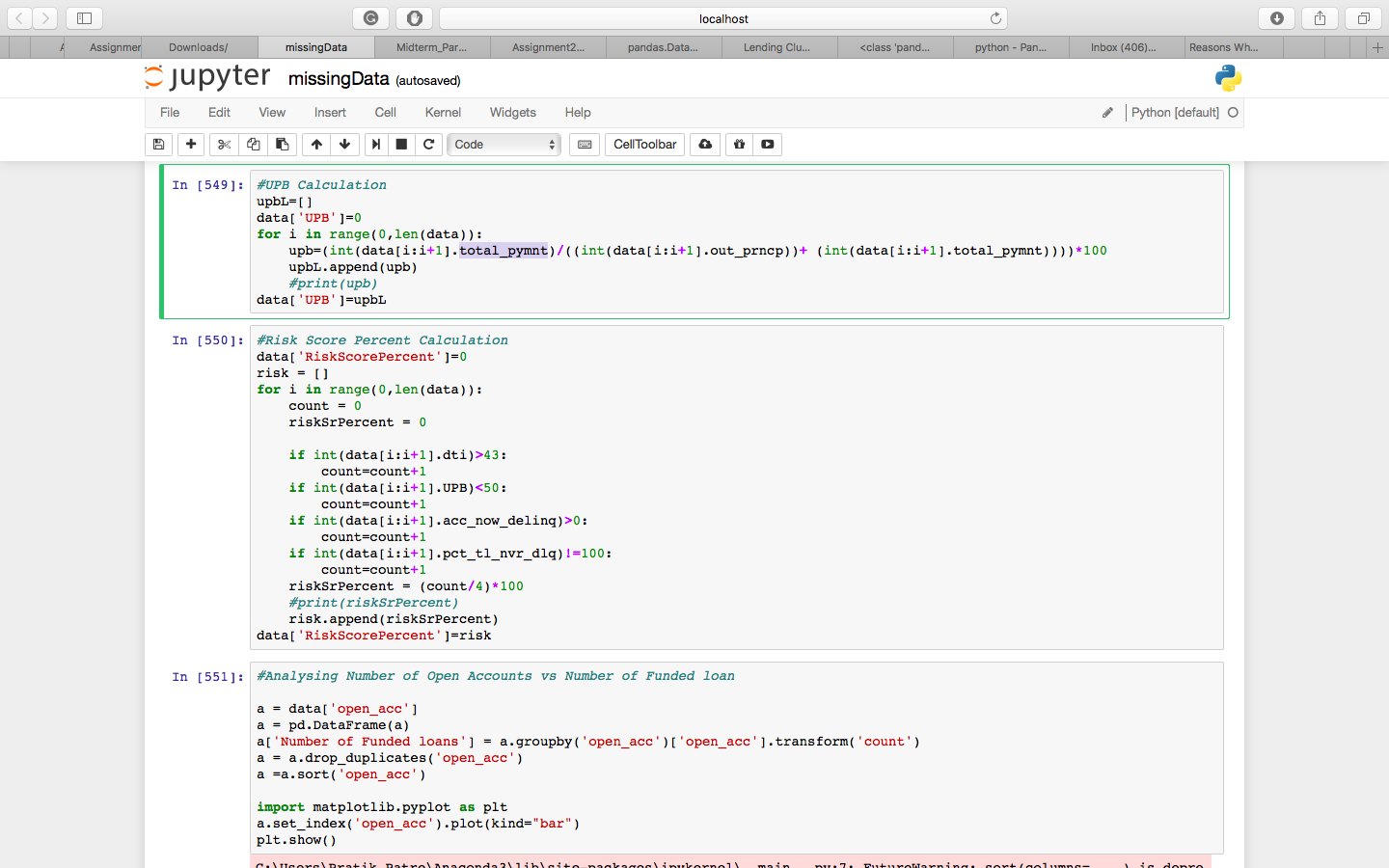
Unpaid Principal Balance is a factor contributing to interest rate. This value is derived by using two columns i.e. paid principal balance and outstanding principal balance. The formula would be:

UPB = unpaid principal/ (remaining balance + paid balance)



**Risk Score Percentage**

The risk score percent decides which all accounts are risky and which are not. We have used UPB which I our derived attribute in deciding this factor. The DTI has to be equal to or less than 43 for qualified loans and so DTI with value greater than 43 is risky. The delinquency value is also considered as it is a contributing factor. If the account has delinquency value of more than 1 then it is considered risky. The percent of trades which were delinquent is also relevant. All these factors have been used to calculate the Risk Score of each account.

****

Step 6: Summarization

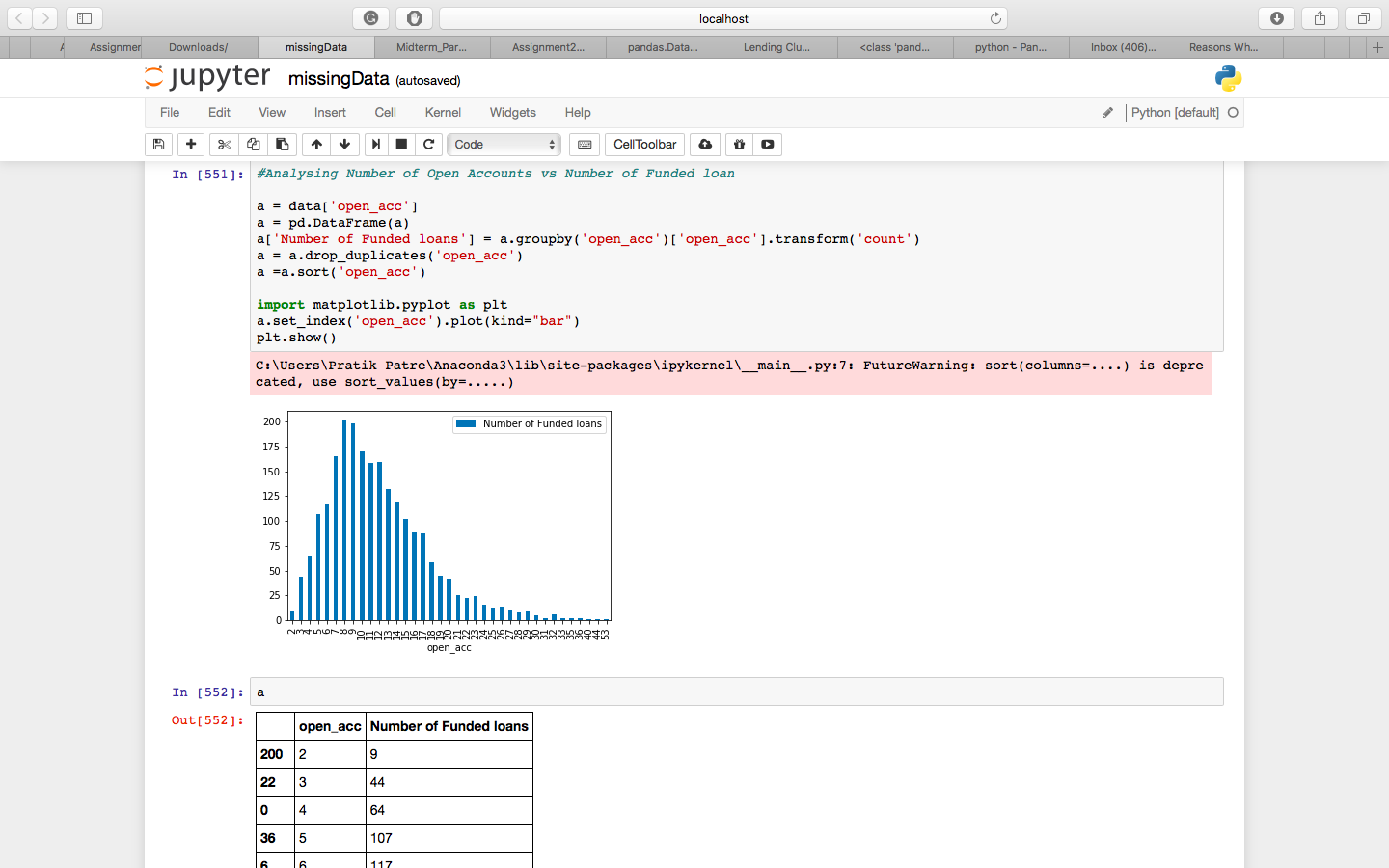
1. We wanted to figure out how effective is the number of open credit lines in the borrower’s credit file on interest rates and net returns.

To figure out which variable is the best to use, we start with some basic diagnostic analyses of the related variables to determine what differences exist.  This includes comparing to see if any are likely inclusive of each other:

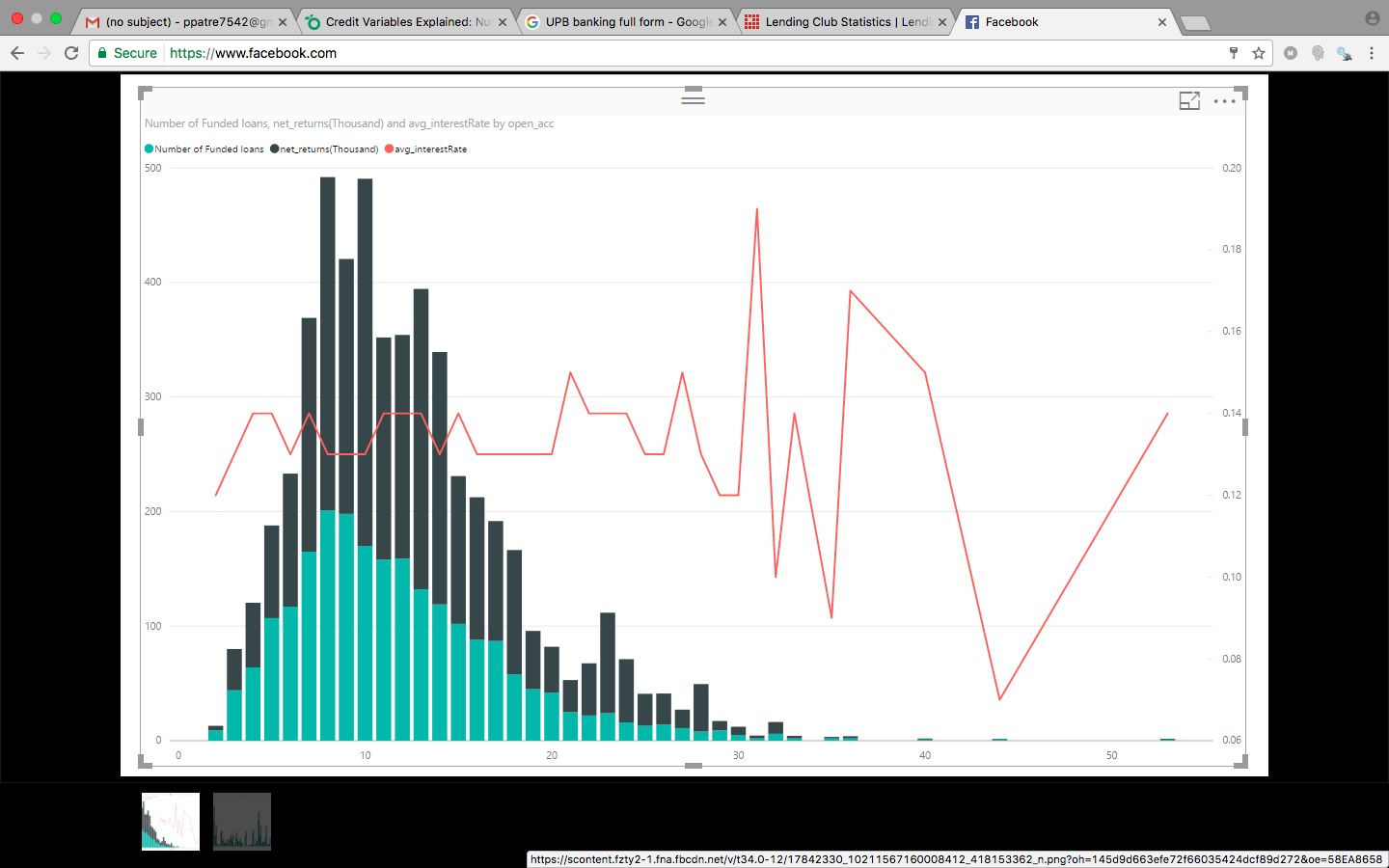
* total\_acc is always larger than or equal to the other variables, confirming the description
* open\_acc is usually smaller than num\_rev\_accts (82% of the time), confirming that num\_rev\_accts probably includes closed accounts

We found that the open\_acc variable is optimal as they give the most accurate picture of the borrower’s current situation.  Looking at how many total (open + closed) accounts is not as useful, since some/all of those accounts may not be available to the borrower, and the numbers can be very different.

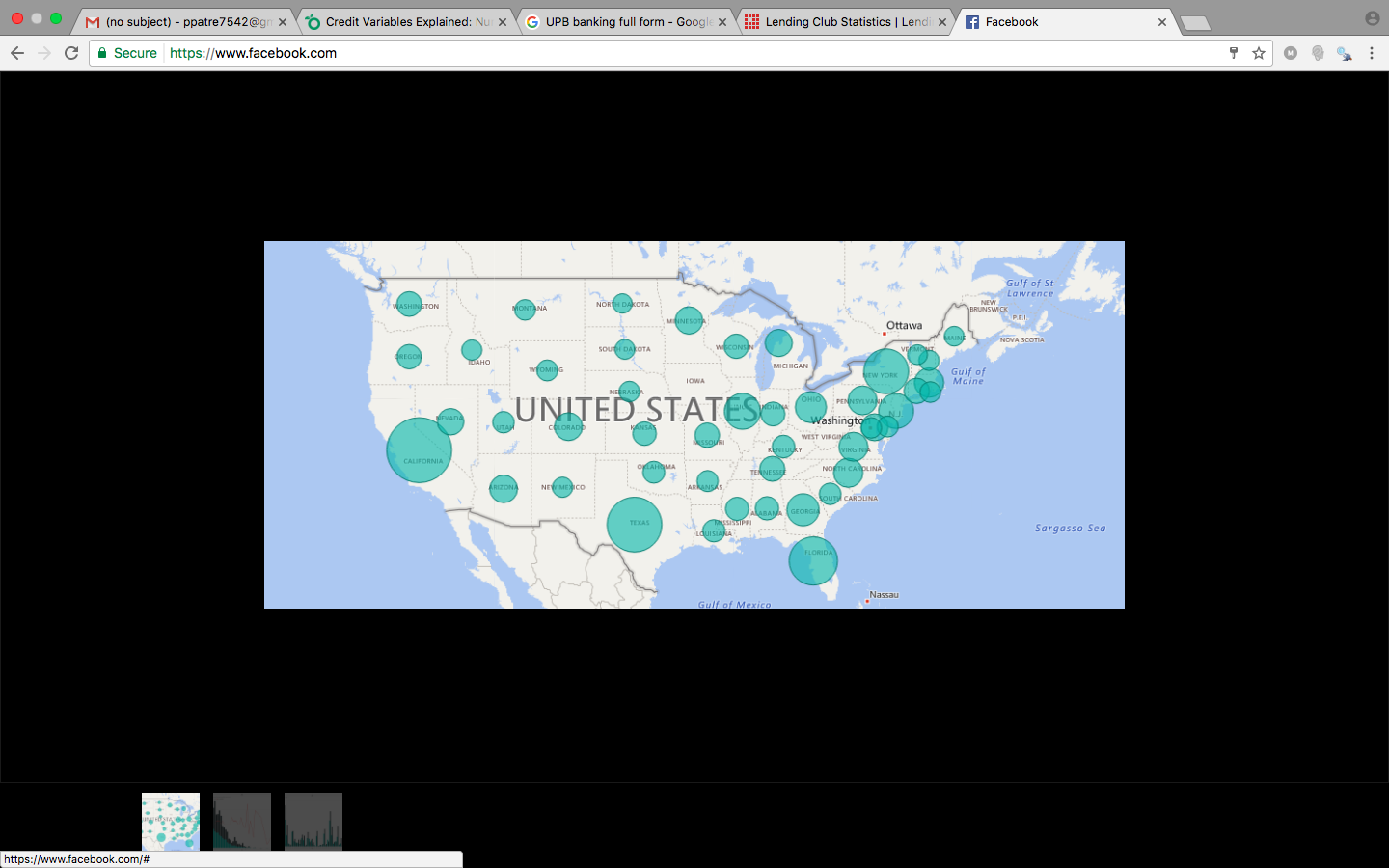
From the below distribution, we see that the bulk of funded loans have < 12 accounts open at a time and also atleast 2 open accounts are necessary.



Now it is important to see the effect on Interest Rate based on net returns. The conclusion derived is that the net returns is less for borrowers with open accounts 2 and also interest rate is high. So it is risky for these account borrowers to have approved loans.



1. Summation of Loan amounts issued across all states in USA. This gives us an idea as to which states are to be focused when loan amounts granted are exceedingly high. We found out that California was having the maximum amount.



1. Number of Accepted Loans and Number of Declined loans.

It is important to analyse how the loans approval rate is fairing for a bank to avoid bankruptcy.



**FILE 2 – Declined Loan Data**

The same process of download and preprocessing will be processed on this file. The next process is of data cleaning there are many assumptions and updates done on the columns.

Clean up of data:

1. Loan Title : Number, date or blank values are changed to ‘other’
2. Risk score : 0 or Space is changed to the minimum value of FICO score (300) if the date is before Nov 5, 2013 or Borrower’s vantage score (300) if the date is after Nov 4, 2013
3. DTI – if value is -1% or 0% then changed to 0 and all other values changed to number
4. Zip code – blank values are changed to ‘000xx’
5. State – blank values are changed to ‘other’
6. Policy code – all 0 values are changed to 1 as mentioned in the description file

New derived columns are added for better analysis of the data

1. Gross Income – It is calculated by using the formula of DTI as it is the ratio of loan amount to gross income
2. Isemployed flag – this is calculated by checking the value of DTI and Employment length. If both are 0 then we assumed it to be not employed and set the value as ‘N’, otherwise ‘Y’.
3. Year – a column added to figure out year of a particular dataset

Summary

We have calculated average risk score, average dti, average Gross income, average amount requested and have observed the trend over the years.

