ADS Team 04

Part1: **Data Download and pre-processing:**

Used logger to capture the execution events. Created the info level logger file and start entering the log in file.

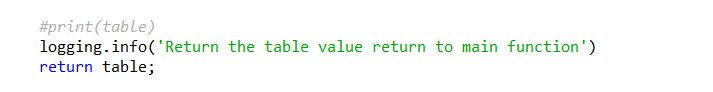
****

Used python payload with required parameters to login into the application. Created the session object to maintain the session through the request. Used same session object for all requests.

Create the second payload “payloadSec” to process the second screen. After processing the second screen, create BeautifulSoup object and find the content based on class=”table1”. Return the table object.

**Error Handling: -** if username and password are not correct, it will not go onto the second screen. Consequently, table value will be none. If table value is none, program will terminate the execution and print the message on console.

****

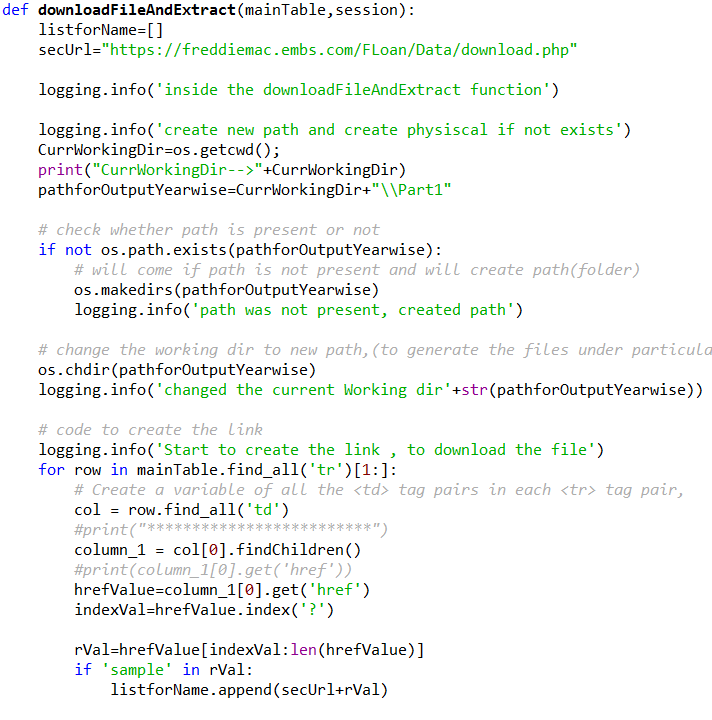


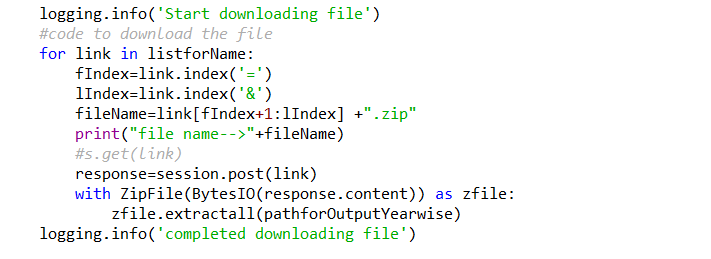
Takes two parameter mainTable and session object. We have already logged in and now will use the same session object to download the files.

Program check for current working dir and add the “part1” to the path and check if the path is present or not. If path is not present than it will create the path and if path is present will do nothing. After that it will set the current working dir to new path.

In the mainTable object find all the tr and iterate through it. Get the td of each row and extract the href value from td. Get the require value by substring the string value. Check the sample name in find string name because need to download only samples files. If sample name is present in string value then it will go inside and create the actual link to download the file.

Make the post request and get the response. Use the ZipFile and BytesIo to convert the response object to file and extractall function to extract the file. File will extract in current working directory.





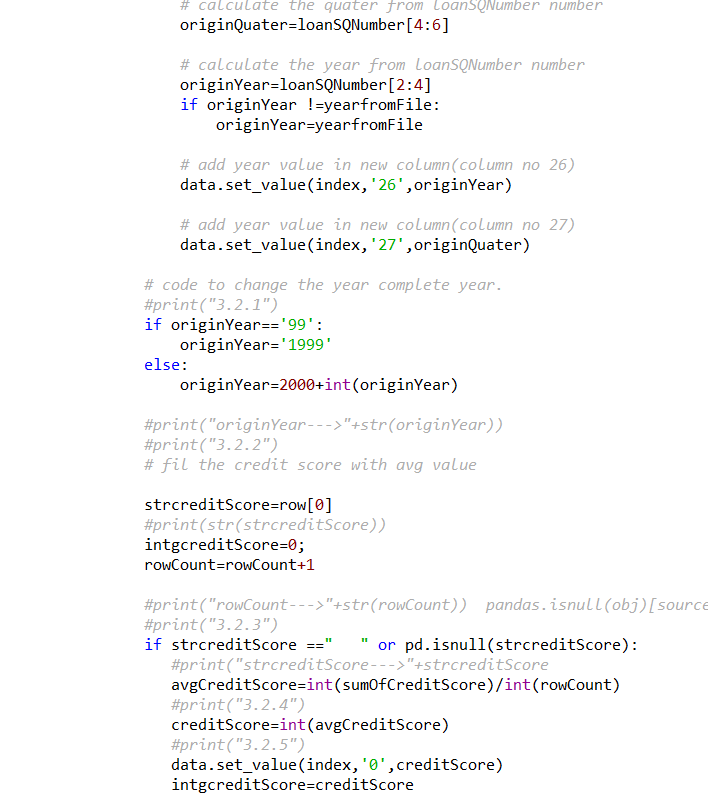
Read the origination files and collect the imortant information. Get the current working directory and iterate over it.check file name contain ‘orig’, if yes then create dataframe using file content. If the loanSequesnceNumber is blank then will drop the complete row. Get the quarter and year information from loansequence number and create two new column in data frame.

If credit score is blank then fill by the average. Check for credit score if credit score it above 700, create new column and enter “Y”. if credit score below than 700 enter “N”

From main data frame get only the quarter one data with the help on new column that we have created.get the avg, Max, low ROI, channel used,product type,state, proprty ued ect values and append in list and create new dataframe.

Calculate the statewise application count. Get the count of first time home buyer over the year.







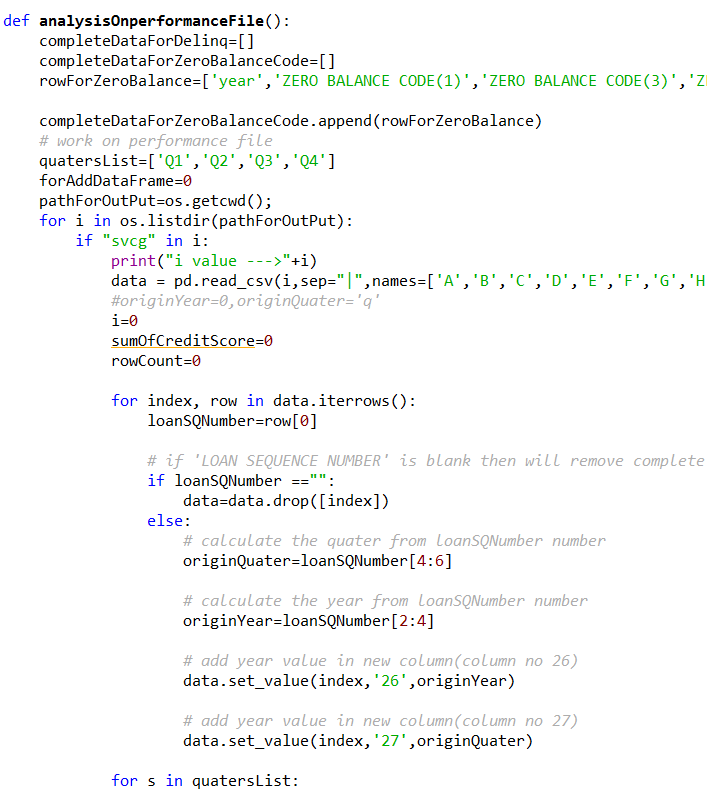




After execution three file generate for quaterAnalysis, stateQuaterAnalysis and for firstTimeHomeBuyer.

Work on performance file and extract the important information. Extract the account information which have Zero balance code as 9 and total number of count based on the zero balance code.

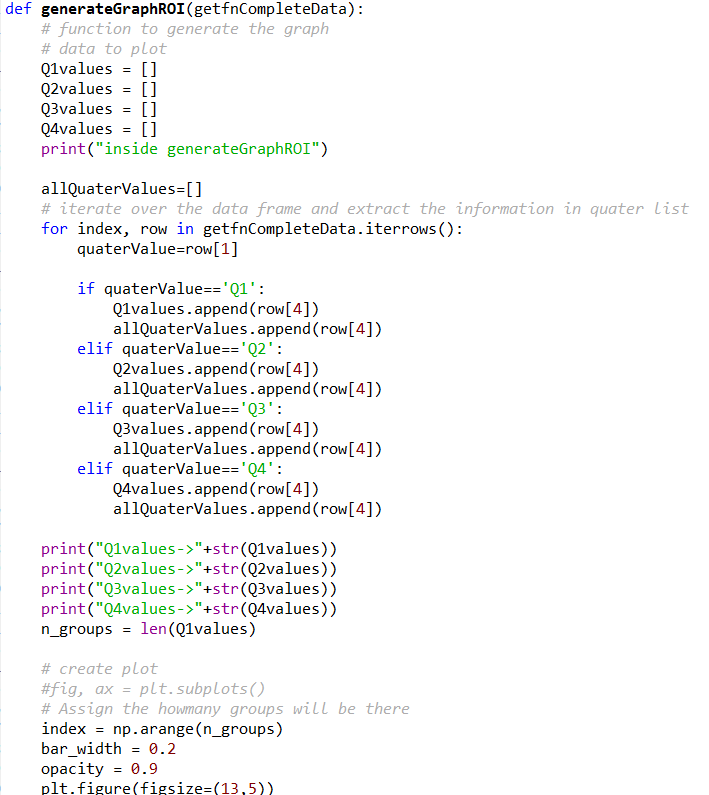
Generate the three file for analysis.

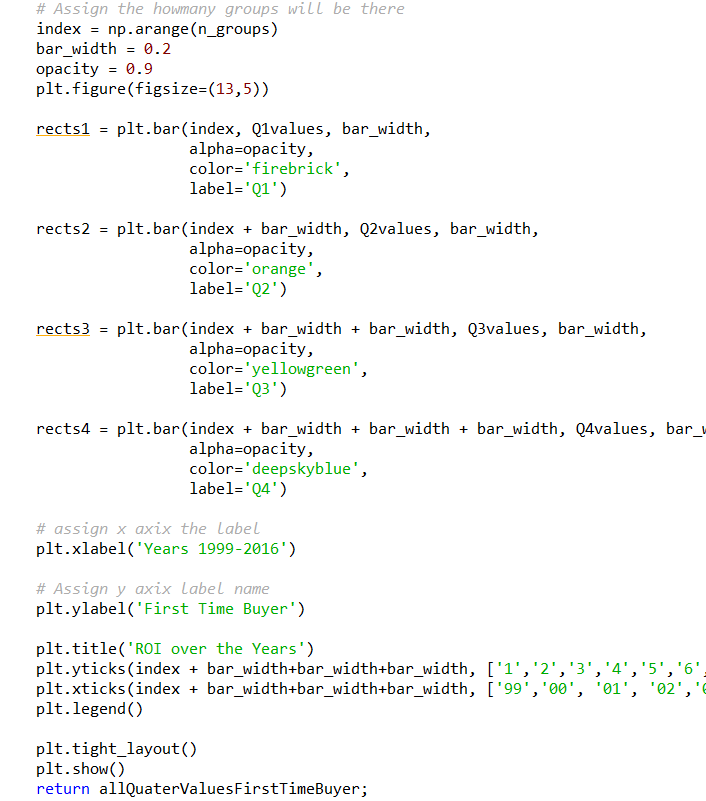




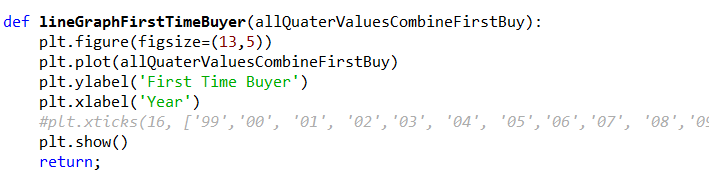


Use matplotlib.pyplot library to generate the graph. Get the completeData as the parameter abd generate the graph for ROI.





Generate the line Graph for ROI over the years.



Generate the Pie chart for occupancy Status.



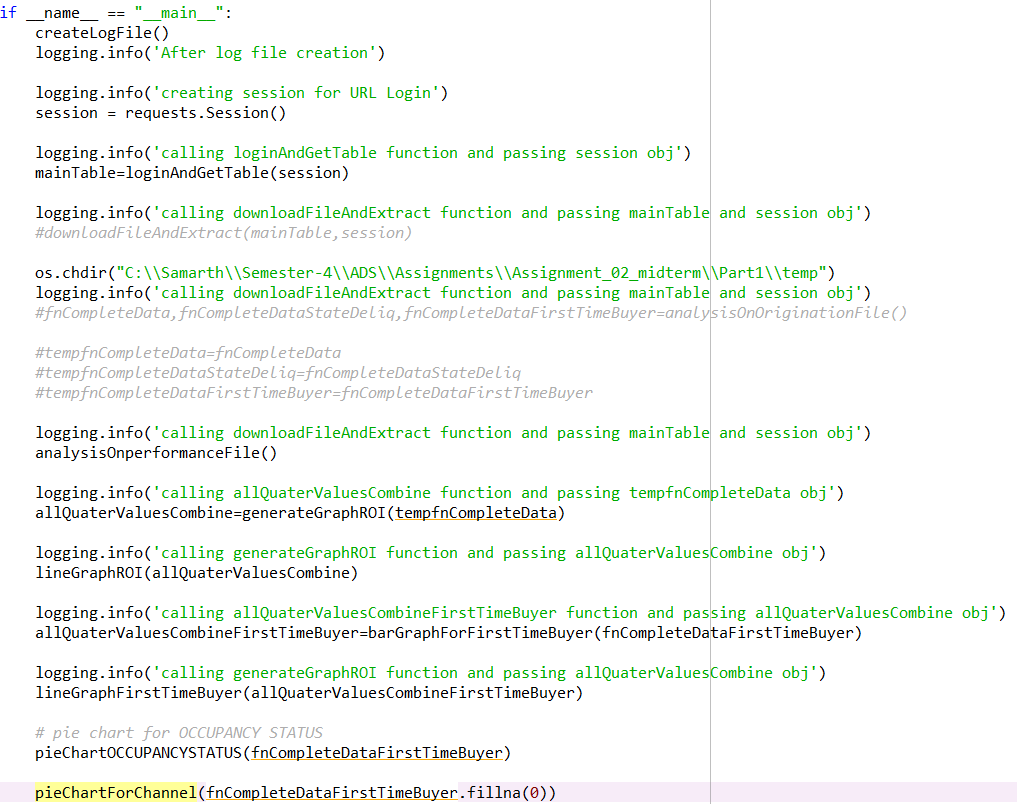


Generate the Pie chart for channel.



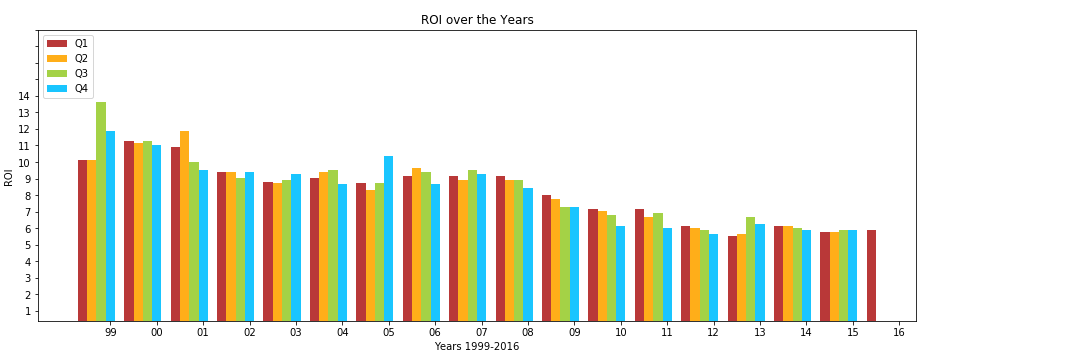


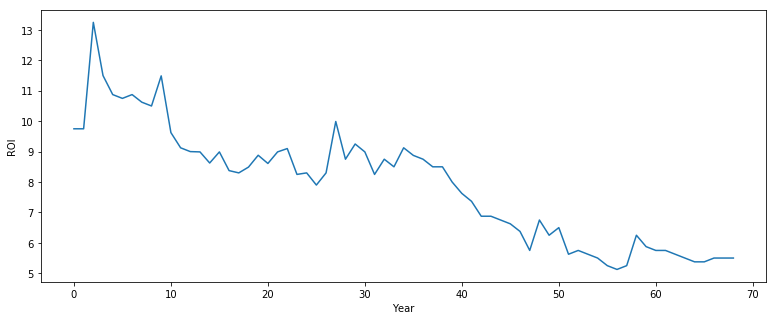
Main function to execution code.



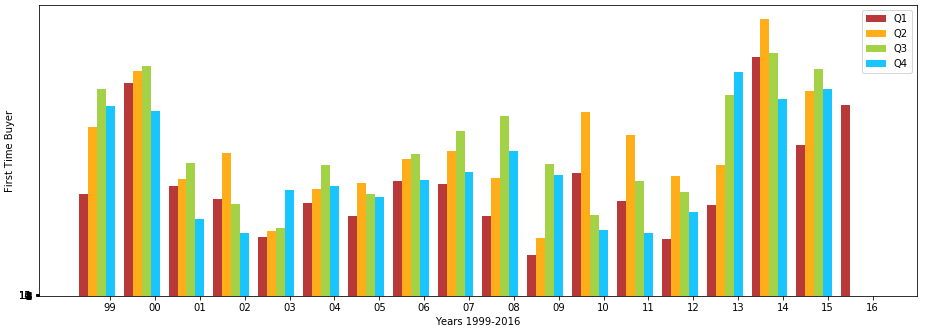
Part 1 :-Write a Jupyter notebook using R/Python to graphically represent different summaries of data. Summarize your findings in this notebook.

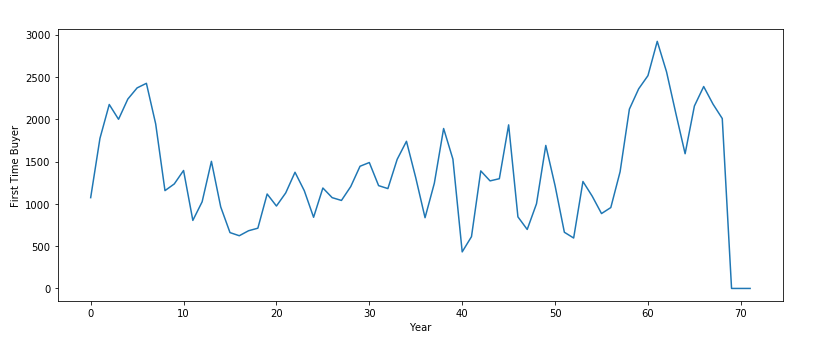
1. ROI Over Years



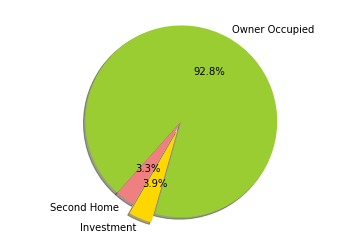


B) First time home buyer.

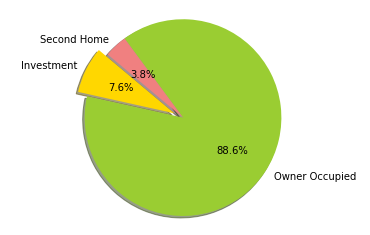




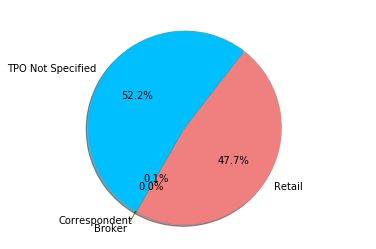
C) OCCUPANCY STATUS for 1999 and 2015



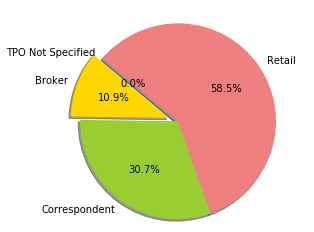
OCCUPANCY STATUS for 2015



D) CHANNEL for 1999



CHANNEL for 2015



b) Create a Tableau based dashboard that analyses the quarterly data from 2005 including summary measures for different attributes, trends over time for different variables. Counts, variability of numerical measures etc. Leverage postal codes and states to indicate location specific information.

**Tableau Public Link**

[https://public.tableau.com/profile/publish/ADSTeam04\_MidTerm/TEAM04\_Midterm#!/publish-confirm](https://public.tableau.com/profile/publish/ADSTeam04_MidTerm/TEAM04_Midterm%23!/publish-confirm)



c) Look at things like interest rate trends, delinquency trends over quarters, location specific insights etc. and other parameters. You have a lot of very interesting variables. Note: You should spend some time analyzing which parameters matter and how things have changed. This is the kind of work of work Data scientists are expected to do. Please take this section seriously!

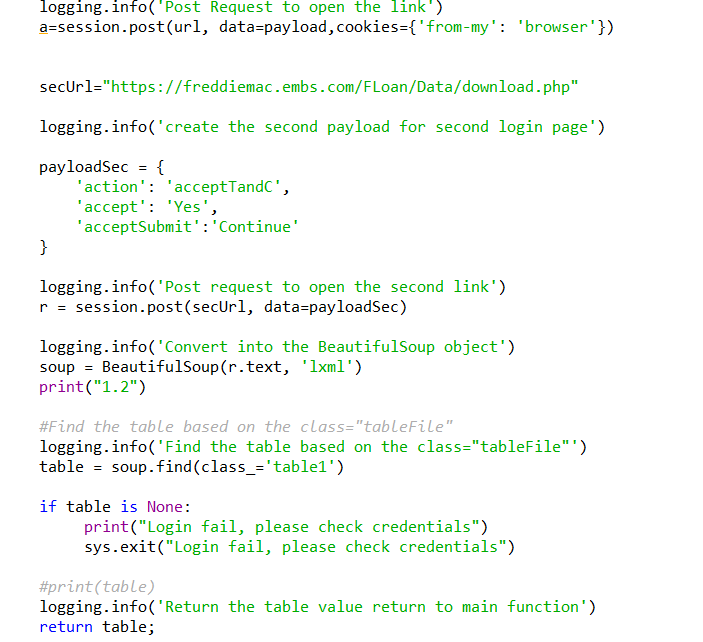


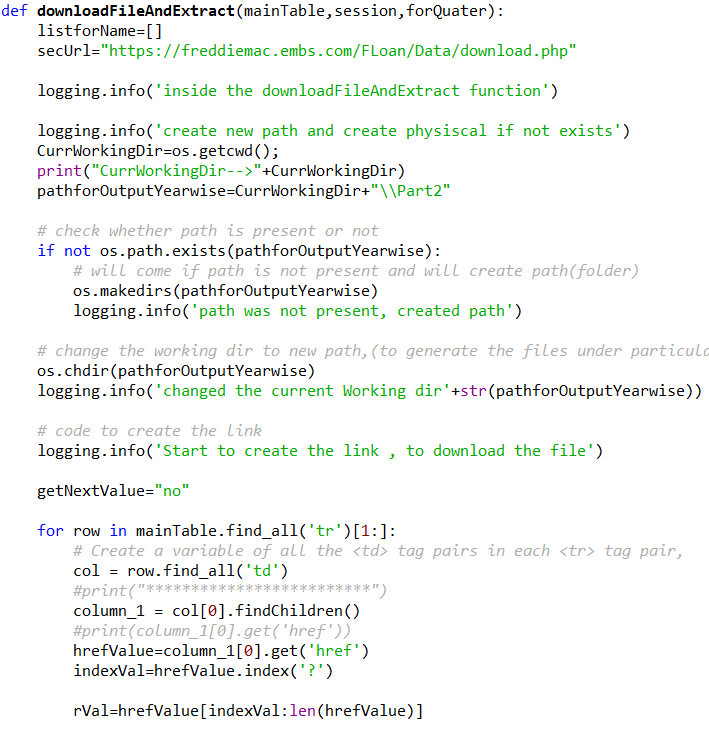
**Part 3: Classification**

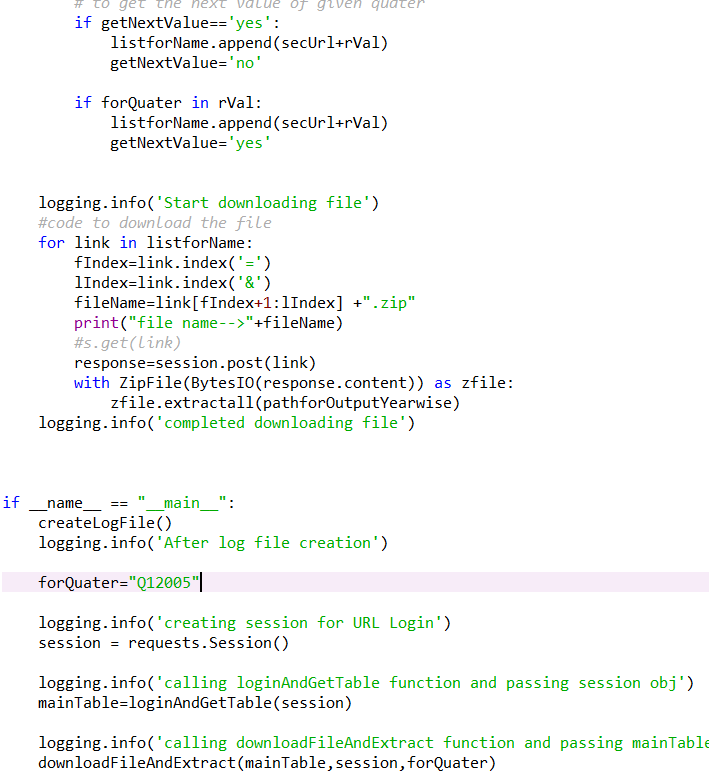
Programmatically downloads Q12005 and Q22005 origination data and pre-processes it.

Code download the files for given quarters.

****

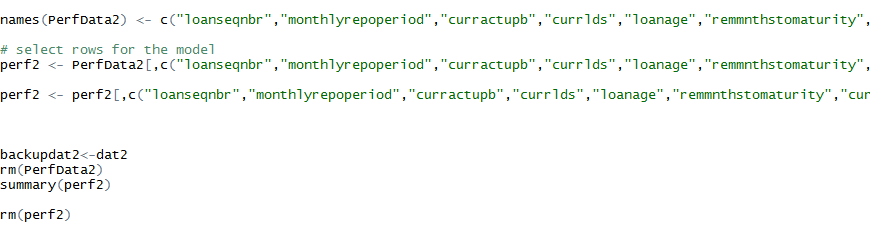
****

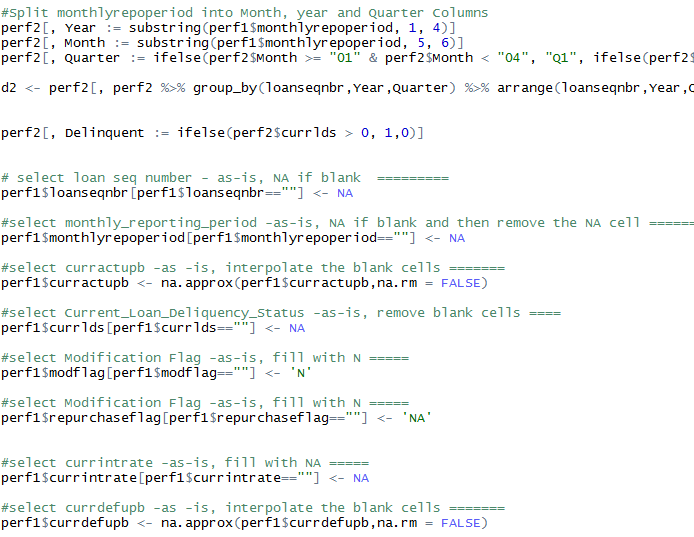
****

****

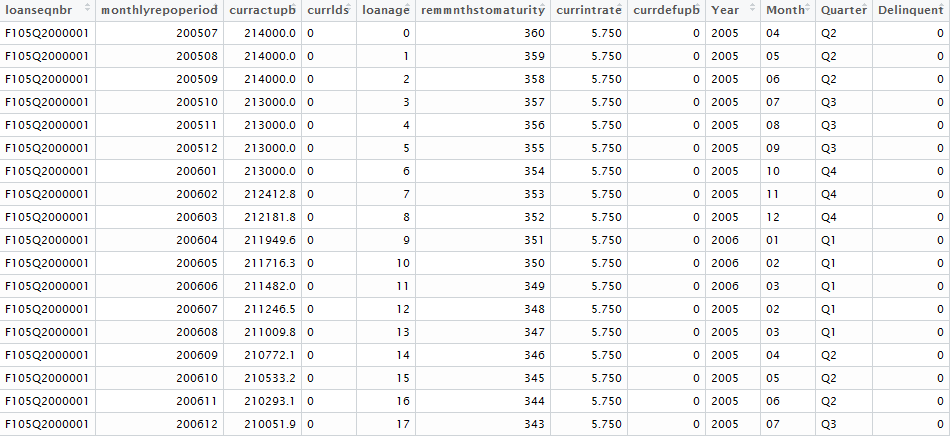
Code for preprocessing: - load the data from file, create dataframe, gives the header name.

Create new column based on column 4, col 4 is > 0, then put one in new column, otherwise put zero.



****

Builds a Logistic regression model for the CURRENT LOAN DELINQUENCY STATUS using Q12005 data as training data (col 4). Note anytime col 4 is > 0, add a new variable as Delinquent. Use this variable as your “Y” variable. IGNORE COL 4 AND DON’T USE IT IN YOUR MODEL.



First Model details:-

Deviance Residuals:

Min 1Q Median 3Q Max

-1.1824 -0.3199 -0.2463 -0.1999 3.2372

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.343e+03 2.880e+01 -46.621 <2e-16 \*\*\*

monthlyrepoperiod 6.666e-03 1.434e-04 46.480 <2e-16 \*\*\*

curractupb 1.793e-06 6.546e-08 27.394 <2e-16 \*\*\*

loanage -3.227e-02 1.145e-03 -28.183 <2e-16 \*\*\*

remmnthstomaturity 4.982e-03 1.497e-04 33.273 <2e-16 \*\*\*

currintrate -1.996e-02 9.581e-03 -2.083 0.0372 \*

currdefupb -1.028e-05 1.011e-06 -10.168 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 360385 on 999999 degrees of freedom

Residual deviance: 336737 on 999993 degrees of freedom

AIC: 336751

Validates against Q22005 data and selects the best Classification model

Code to validate the Q12005 and Q2005.

library(forecast)

full.pred = predict(fit1, perf2,type='response')

require(forecast)

accuracy(full.pred, perf1$Delinquent)

Out put:-

ME RMSE MAE MPE MAPE

Test set -0.005021074 0.2088053 0.08852242 -Inf Inf

Computes ROC curve and Confusion matrices for training and testing datasets.

Confusion matrices for training and testing datasets.

pred

0 1

0 953306 12

1 46592 90

