Advanced Data Science & Architecture

Midterm Project

* *Under the guidance of Sri Krishnamurthy*

Compiled By,

Mohit Mittal,

Sneha Ravikumar,

Taj Poovaiah

Contents

[Downloading Data: 7](#_Toc477551294)

[Origination File Observation and Cleaning 8](#_Toc477551295)

[Credit Score: Deleted the rows that had missing credit score 8](#_Toc477551296)

[FIRST PAYMENT DATE: No missing values in sample files 8](#_Toc477551297)

[FIRST TIME HOMEBUYER FLAG: 8](#_Toc477551298)

[MATURITY DATE: 8](#_Toc477551299)

[METROPOLITAN STATISTICAL AREA(MSA) OR METROPOLITAN DIVISION: 9](#_Toc477551300)

[MORTGAGE INSURANCE PERCENTAGE (MI%): 9](#_Toc477551301)

[NUMBER OF UNITS: 10](#_Toc477551302)

[OCCUPANCY STATUS: 10](#_Toc477551303)

[ORIGINAL COMBINED LOAN-TO-VALUE(CLTV): 10](#_Toc477551304)

[ORIGINAL INTEREST RATE: 12](#_Toc477551305)

[CHANNEL: 12](#_Toc477551306)

[PREPAYMENT PENALTY MORTGAGE (PPM) FLAG: 12](#_Toc477551307)

[PRODUCT TYPE: 13](#_Toc477551308)

[PROPERTY STATE: 13](#_Toc477551309)

[PROPERTY TYPE: 13](#_Toc477551310)

[POSTAL CODE: 14](#_Toc477551311)

[LOAN SEQUQENCE NUMBER: 15](#_Toc477551312)

[LOAN PURPOSE: 15](#_Toc477551313)

[ORIGINAL LOAN TERM: 15](#_Toc477551314)

[NUMBER OF BORROWERS: 15](#_Toc477551315)

[SELLER NAME: 16](#_Toc477551316)

[SERVICES NAME: 16](#_Toc477551317)

[SUPER CONFORMING FLAG: 16](#_Toc477551318)

[PERFORMANCE FILE 16](#_Toc477551319)

[LOAN SEQUENCE NUMBER: 16](#_Toc477551320)

[MONTHLY REPORTING PERIOD: 17](#_Toc477551321)

[CURRENT ACTUAL UPB: 17](#_Toc477551322)

[CURRENT LOAN DELINQUENCY STATUS: 17](#_Toc477551323)

[LOAN AGE: 17](#_Toc477551324)

[REMAINING MONTHS TO LEGAL MATURITY: 17](#_Toc477551325)

[REPURCHASE FLAG: 17](#_Toc477551326)

[MODIFICATION FLAG: 17](#_Toc477551327)

[ZERO BALANCE CODE: 17](#_Toc477551328)

[ZERO BALANCE EFFECTIVE DATE: 17](#_Toc477551329)

[CURRENT INTEREST RATE: 17](#_Toc477551330)

[DUE DATE OF LAST PAID INSTALLMENT: 17](#_Toc477551331)

[Replacing missing values with 0 for the following columns 17](#_Toc477551332)

[Exploratory Data analysis 18](#_Toc477551333)

[SUMMARIES OF SAMPLE ORIGINAL FILES- 1999 18](#_Toc477551334)

[Sum of Original UPB PER YEAR (1999) 18](#_Toc477551335)

[OBSERVATION 18](#_Toc477551336)

[Sum of Original UPB per QUARTER 18](#_Toc477551337)

[Sum of Original UPB per YEAR PER Quarter 19](#_Toc477551338)

[Quarter 1 19](#_Toc477551339)

[Quarter 2 19](#_Toc477551340)

[Quarter 3 20](#_Toc477551341)

[Quarter 4 20](#_Toc477551342)

[AVERAGE of Original UPB per year 21](#_Toc477551343)

[Average of Original UPB per year per Quarter 22](#_Toc477551344)

[Average of Original UPB per year per Quarter 22](#_Toc477551345)

[QUARTER 1 22](#_Toc477551346)

[QUARTER 2 23](#_Toc477551347)

[QUARTER 3 24](#_Toc477551348)

[QUARTER 4 24](#_Toc477551349)

[Average of Credit Score per year 24](#_Toc477551350)

[OBSERVATION 25](#_Toc477551351)

[Average of Credit Score per Quarter 25](#_Toc477551352)

[QUARTER 1 25](#_Toc477551353)

[QUARTER 2 26](#_Toc477551354)

[QUARTER 3 27](#_Toc477551355)

[QUARTER 4 27](#_Toc477551356)

[Average of Credit Score per year per Quarter 28](#_Toc477551357)

[Average of CLTV per year 28](#_Toc477551358)

[Average of CLTV per Quarter 28](#_Toc477551359)

[Average of CLTV per year per Quarter 29](#_Toc477551360)

[QUARTER 1 29](#_Toc477551361)

[QUARTER 2 29](#_Toc477551362)

[QUARTER 3 30](#_Toc477551363)

[QUARTER 4 30](#_Toc477551364)

[Average of LTV per Year 30](#_Toc477551365)

[Average of LTV per Quarter 31](#_Toc477551366)

[Average of LTV per year per Quarter 31](#_Toc477551367)

[QUARTER 1 31](#_Toc477551368)

[QUARTER 2 32](#_Toc477551369)

[QUARTER 3 33](#_Toc477551370)

[QUARTER 4 33](#_Toc477551371)

[Average of Interest Rate per Year 34](#_Toc477551372)

[OBSERVATION 34](#_Toc477551373)

[Average of Interest Rate per Quarter 34](#_Toc477551374)

[Average of Interest Rate per year per Quarter 34](#_Toc477551375)

[QUARTER 1 35](#_Toc477551376)

[QUARTER 2 35](#_Toc477551377)

[QUARTER 3 36](#_Toc477551378)

[QUARTER 4 36](#_Toc477551379)

[Count of Loans with First Time Home Buyer equal to "Y", Occupancy equal to "I" or "S" and Loan Purpose equal to "C" and "N" 37](#_Toc477551380)

[Count of Loans with MSA flag equal to "YES" 37](#_Toc477551381)

[Count of Loans with MSA flag equal to "NO" 37](#_Toc477551382)

[Average Original UPB where MSA flag equal to "YES" 38](#_Toc477551383)

[Average Original UPB where MSA flag equal to "YES" 38](#_Toc477551384)

[Average Credit Score where MSA flag equal to "YES" 38](#_Toc477551385)

[Average Credit Score where MSA flag equal to "NO" 38](#_Toc477551386)

[Average Interest Rate where MSA flag equal to "YES" 38](#_Toc477551387)

[Average Interest Rate where MSA flag equal to "NO" 38](#_Toc477551388)

[Count of Loans where PPM flag equal to "Y" 38](#_Toc477551389)

[SUMMARIES OF SAMPLE PERFORMANCE FILES- 2016 38](#_Toc477551390)

[DISTINCT COUNT OF LOAN 38](#_Toc477551391)

[COUNT OF LOANS WITH CURRENT UPB EQUAL TO 0, AND ZERO BALANCE CODE EQUAL TO 1 OR 6 39](#_Toc477551392)

[Count of ACTUAL LOSS CALCULATION with Current UPB equal to 0, and Zero balance code equal to 9 40](#_Toc477551393)

[0BSERVATION 40](#_Toc477551394)

[COUNT OF Loans with Current UPB not equal to 0 40](#_Toc477551395)

[0BSERVATION 41](#_Toc477551396)

[Count of Loans with DELINQUENCY STATUS >= 5 41](#_Toc477551397)

[Count of Loans with Current UPB not equal to 0 and DELINQUENCY STATUS >= 0 41](#_Toc477551398)

[0BSERVATION 41](#_Toc477551399)

[Count of Modification flag grouped by Loan Sequence Number 41](#_Toc477551400)

[**OBSERVATIONS** 44](#_Toc477551401)

[Downloading Data(Part 2): 45](#_Toc477551402)

[Origination File Observation and Cleaning 46](#_Toc477551403)

[FIRST PAYMENT DATE: 46](#_Toc477551404)

[FIRST TIME HOMEBUYER FLAG: 46](#_Toc477551405)

[MATURITY DATE: 47](#_Toc477551406)

[METROPOLITAN STATISTICAL AREA(MSA) OR METROPOLITAN DIVISION: 47](#_Toc477551407)

[MORTGAGE INSURANCE PERCENTAGE (MI%): 47](#_Toc477551408)

[NUMBER OF UNITS: 48](#_Toc477551409)

[OCCUPANCY STATUS: 48](#_Toc477551410)

[ORIGINAL COMBINED LOAN-TO-VALUE(CLTV): 48](#_Toc477551411)

[ORIGINAL DEBT-TO-INCOME (DTI) RATIO: 49](#_Toc477551412)

[ORIGINAL UPB: 49](#_Toc477551413)

[ORIGINAL LOAN-TO-VALUE: 49](#_Toc477551414)

[ORIGINAL INTEREST RATE: 50](#_Toc477551415)

[CHANNEL: 50](#_Toc477551416)

[PREPAYMENT PENALTY MORTGAGE (PPM) FLAG: 50](#_Toc477551417)

[PROPERTY TYPE: 51](#_Toc477551418)

[POSTAL CODE: 52](#_Toc477551419)

[LOAN SEQUQENCE NUMBER: 53](#_Toc477551420)

[LOAN PURPOSE: 53](#_Toc477551421)

[ORIGINAL LOAN TERM: 53](#_Toc477551422)

[NUMBER OF BORROWERS: 53](#_Toc477551423)

[SELLER NAME: 54](#_Toc477551424)

[SERVICES NAME: 54](#_Toc477551425)

[SUPER CONFORMING FLAG: 54](#_Toc477551426)

[PERFORMANCE FILE 54](#_Toc477551427)

[CLASSIFICATION (LOGISTIC REGRESSION) 55](#_Toc477551428)

[summary of the logistic regression 55](#_Toc477551429)

[CONFUSION MATRIX for LOGISTIC REGRESSION 56](#_Toc477551430)

[prediction 56](#_Toc477551431)

[ROC CURVE 56](#_Toc477551432)

[NEURAL NETWORK 57](#_Toc477551433)

[CONFUSION MATRIX for NEURAL NETWORK 58](#_Toc477551434)

[PREDICTION 58](#_Toc477551435)

# STEPS TO RUN THE CODES

Part1:

1. Open the terminal in the directory "1\_Final\_Code"

2. To run the program enter following command without quotes:

"python part1.py Summarize\_data --local-scheduler"

Pre-requisites to run the program:

- python3.x

- pip

- python libraries: luigi, beautifulsoup, mechanicalsoup, glob, pandas, numpy

The program will ask you for the username and password for freddiemac.embs.com website to download the data.

3. Enter the username and password and it will run the tasks to download the data, clean the data and summarize it

4. The downloaded files can be found in the "1\_Final\_Code/downloads" directory, the cleaned files can be found in the "1\_Final\_Code/cleaned" directory and the summaries can be found in the "1\_Final\_Code/summary" directory

5. The python notebook for the summary can be found in the "1\_Final\_Code" directory with the name "New\_Summary\_Performance". The tableu files can be found in the directory ""1\_Final\_Code/TableauFiles"

Part2:

1. Open the terminal in the directory "1\_Final\_Code"

2. To run the program enter following command without quotes:

"python part2.py Build\_prediction\_model --local-scheduler"

Pre-requisites to run the program:

- python3.x

- pip

- python libraries: luigi, beautifulsoup, mechanicalsoup, glob, pandas, numpy

The program will ask you for the username and password for freddiemac.embs.com website to download the data.

3. Enter the username and password. Enter the year and quarter you want to run the prediction model for. it will run the tasks to download the data, clean the data and summarize it.

4. The downloaded files can be found in the "1\_Final\_Code/downloads" directory and the cleaned files can be found in the "1\_Final\_Code/cleaned" directory.

5. Run the "Classification\_Logistic\_Regression.R" in RStudio to run the logistic regression for Delinquent, "neuralnet.R" in RStudio to run the neural network for Delinquent.

**Programming Language used : Python**

**Workflow Manager User: Luigi**

**Tasks**

1. **Downloading Data**
2. **Clean Origination Data**
3. **Clean Performance Data**
4. **Summarizing Origination Data**
5. **Summarizing Performance Data**

# Downloading Data:

File Location : Classes/Part1/Download\_sf\_loan.py

Task Requires no prior tasks to be completed.

Output of the task are all the sample origination and performance files.

Process:

* Asking user for username and password.
* Creating a browser agent (using the mechanicalsoup library) to store and pass the cookies
* Logging in with the user’s credentials
* Checking if the user is succesfuly logged in or not.
* Landing to the page that contains the list of files and download links
* Putting the table of files in a dataframe
* Iterating through the rows in dataframe for the links that contain sample files and downloading them to a newly created (if it doesn’t already exist) “Downloads” directory
* The program also checks if the files are already present in the “Downloads” directory. It skips the downloading if the file already exists.
* Unzipping the downloaded file.

# Origination File Observation and Cleaning

## Credit Score: Deleted the rows that had missing credit score

* 1. Cannot replace missing values as it is explicitly specified that credit score can be either less than 301 or greater than 850.
  2. Number of such instances is very less (0.002% in 2016 to 1.242% in 2000)
  3. Removing the rows that have blank values and nulls for credit score.

|  |  |  |
| --- | --- | --- |
| CREDIT SCORE | COUNT OF BLANKS | YEAR |
|  | 362 | 1999 |
|  | 621 | 2000 |
|  | 274 | 2001 |
|  | 201 | 2002 |
|  | 33 | 2003 |
|  | 42 | 2004 |
|  | 24 | 2005 |
|  | 39 | 2006 |
|  | 29 | 2007 |
|  | 29 | 2008 |
|  | 1 | 2009 |
|  | 1 | 2011 |
|  | 3 | 2013 |
|  | 2 | 2014 |
|  | 1 | 2016 |

## FIRST PAYMENT DATE: No missing values in sample files

## FIRST TIME HOMEBUYER FLAG:

* 1. If blank it can be replaced by NA if Occupancy Status is either “I” or “S” (Investement property or Second Home)
  2. If blank it can replaced by NA if Loan Purpose is either “C” or “N” (Refinance)
  3. If blank, then replace it with NA
  4. Created three columns for – First Time HomeBuyer Flag YES (1,0) , NO(1,0) and NA(1,0)

## MATURITY DATE:

* 1. No missing values in the sample files
  2. Splitting Maturity year and month

## METROPOLITAN STATISTICAL AREA(MSA) OR METROPOLITAN DIVISION:

* 1. Replaced missing values with zero.
  2. Derived a new column for Metropolitan Area Flag, that had values in it
  3. Future Scope: Compare the values of zip codes, if the zip code belongs to a MSA or MD, then map the msa or md code in the data.

|  |  |
| --- | --- |
| YEAR | COUNT OF BLANKS |
| 1999 | 7640 |
| 2000 | 7542 |
| 2001 | 6978 |
| 2002 | 7309 |
| 2003 | 7182 |
| 2004 | 7844 |
| 2005 | 7913 |
| 2006 | 8209 |
| 2007 | 8671 |
| 2008 | 7729 |
| 2009 | 7528 |
| 2010 | 7022 |
| 2011 | 6944 |
| 2012 | 6593 |
| 2013 | 5475 |
| 2014 | 5030 |
| 2015 | 4845 |
| 2016 | 1184 |

## MORTGAGE INSURANCE PERCENTAGE (MI%):

|  |  |  |  |
| --- | --- | --- | --- |
| MORTGAGE INSURANCE PERCENTAGE (MI %) | COUNT | YEAR | Percentage |
|  | 9026 | 1999 | 18.052 |
| 0 | 21885 | 1999 | 43.77 |
|  | 44 | 2000 | 0.088 |
| 0 | 32764 | 2000 | 65.528 |
|  | 58 | 2001 | 0.116 |
| 0 | 36990 | 2001 | 73.98 |
|  | 11 | 2002 | 0.022 |
| 0 | 38304 | 2002 | 76.608 |
|  | 10 | 2003 | 0.02 |
| 0 | 40083 | 2003 | 80.166 |
|  | 9 | 2004 | 0.018 |
| 0 | 40437 | 2004 | 80.874 |
|  | 57 | 2005 | 0.114 |
| 0 | 43136 | 2005 | 86.272 |
| 0 | 43086 | 2006 | 86.172 |
| 0 | 39839 | 2007 | 79.678 |
| 0 | 40958 | 2008 | 81.916 |
| 0 | 46460 | 2009 | 92.92 |
| 0 | 46266 | 2010 | 92.532 |
| 0 | 44985 | 2011 | 89.97 |
| 0 | 43711 | 2012 | 87.422 |
| 0 | 40459 | 2013 | 80.918 |
| 0 | 36478 | 2014 | 72.956 |
| 0 | 37309 | 2015 | 74.618 |
| 0 | 9481 | 2016 | 18.962 |

* 1. Zero means No Mortgage insurance
  2. Blanks Means either less than 1% or greater than 55%, so the replacement cannot be generalized in this case. Also, such cases are ~18% in 1999 and ~0.01% in until 2005 and 0 in the later years.
  3. Deriving a new column for mortgage insurance flag is done, where the value is kept No if MI% is zero, otherwise it is made Yes

## NUMBER OF UNITS:

|  |  |  |
| --- | --- | --- |
|  | 1 | 2000 |
|  | 7 | 2004 |

* 1. No missing values for most sample files. Only 1 in the year 2000 and 7 cases in 2004 where number of units is missing
  2. Replaced it with the mode OR Discard the row

## OCCUPANCY STATUS:

* 1. No missing values in the sample files.
  2. Handled the missing value by replacing it by mode or discarding the rows

## ORIGINAL COMBINED LOAN-TO-VALUE(CLTV):

|  |  |  |
| --- | --- | --- |
| ORIGINAL COMBINED LOAN-TO-VALUE (CLTV) | COUNT | Year |
|  | 0 | 1999 |
|  | 3 | 2000 |
|  | 2 | 2001 |
|  | 4 | 2002 |
|  | 3 | 2003 |
|  | 3 | 2004 |
|  | 6 | 2005 |
|  | 1 | 2006 |
|  | 2 | 2007 |
|  | 0 | 2008 |
|  | 0 | 2009 |
|  | 1 | 2010 |
|  | 0 | 2011 |
|  | 2 | 2012 |
|  | 2 | 2013 |
|  | 1 | 2014 |
|  | 2 | 2015 |
|  | 0 | 2016 |

* 1. ~0.01% missing values in the sample files.
  2. If the LTV is less than 80 or greater than 200 or unknown, then this column is unknown. Also if CLTV is less than LTV then, CLTV is set to unknown.
  3. This value is dependent on each individual case, so may not be replaced by mean, median or mode.

1. ORIGINAL DEBT-TO-INCOME (DTI) RATIO:
   1. Ratio greater than 65% are represented as spaces. We replaced it by 70.
   2. Unknowns are represented by null, which we replaced by the median.
2. ORIGINAL UPB:
   1. No missing values in the sample files
   2. If value is missing then discard the rows.
3. ORIGINAL LOAN-TO-VALUE:
   1. Ratios below 6% and greater than 105% are unknown.

|  |  |  |
| --- | --- | --- |
| ORIGINAL LOAN-TO-VALUE (LTV) | COUNT | YEAR |
|  | 0 | 1999 |
|  | 2 | 2000 |
|  | 1 | 2001 |
|  | 1 | 2002 |
|  | 3 | 2003 |
|  | 3 | 2004 |
|  | 6 | 2005 |
|  | 1 | 2006 |
|  | 2 | 2007 |
|  | 0 | 2008 |
|  | 0 | 2009 |
|  | 1 | 2010 |
|  | 0 | 2011 |
|  | 2 | 2012 |
|  | 2 | 2013 |
|  | 1 | 2014 |
|  | 2 | 2015 |
|  | 0 | 2016 |

* 1. Close to zero percent of such occurrence. But, replacing of the values with mean/median cannot be justified as it is specifically said that these values are either less than 6 or greater than 105. So, discarding such rows.

## ORIGINAL INTEREST RATE:

* 1. No missing values
  2. If value is missing then replace by median

## CHANNEL:

* 1. No missing values in sample files
  2. If values are missing then replace by mode

## PREPAYMENT PENALTY MORTGAGE (PPM) FLAG:

|  |  |  |
| --- | --- | --- |
| PREPAYMENT PENALTY MORTGAGE (PPM) FLAG | COUNT | YEAR |
|  | 1247 | 1999 |
|  | 236 | 2000 |
|  | 122 | 2001 |
|  | 171 | 2002 |
|  | 198 | 2003 |
|  | 73 | 2004 |
|  | 49 | 2005 |
|  | 65 | 2006 |
|  | 113 | 2007 |
|  | 1039 | 2008 |
|  | 317 | 2009 |
|  | 336 | 2010 |
|  | 580 | 2011 |
|  | 39 | 2012 |
|  | 4 | 2013 |
|  | 11 | 2014 |
|  | 41 | 2015 |
|  | 7 | 2016 |

* 1. Most number of blanks (unknown) in the year 1999 -> 2.49%, 2008 -> 2.078%

|  |  |
| --- | --- |
| **1999** | **48753** |
| N | 48491 |
| Y | 262 |
| **2000** | **49764** |
| N | 49737 |
| Y | 27 |
| **2001** | **49878** |
| N | 49867 |
| Y | 11 |
| **2002** | **49829** |
| N | 49784 |
| Y | 45 |
| **2003** | **49802** |
| N | 49652 |
| Y | 150 |
| **2004** | **49927** |
| N | 49752 |
| Y | 175 |

* 1. Maximum are “N” throughout the years. 97.5% in 1999, 99.5% in 2000…
  2. We are replacing unknown(blanks) values by mode as it wouldn’t affect the distribution.

## PRODUCT TYPE:

* 1. No missing values found in the observations
  2. If there are any missing values, then it is replaced with “FRM”

## PROPERTY STATE:

* 1. No missing values found in the observations
  2. If there are any missing values, then it is replaced with “Unknown”

## PROPERTY TYPE:

|  |  |  |
| --- | --- | --- |
| PROPERTY TYPE | COUNT | YEAR |
|  | 8 | 2000 |
|  | 11 | 2001 |
|  | 3 | 2002 |
|  | 14 | 2004 |

|  |  |  |
| --- | --- | --- |
| PROPERTY TYPE | COUNT | YEAR |
|  | 8 | 2000 |
| CO | 4090 | 2000 |
| CP | 74 | 2000 |
| LH | 15 | 2000 |
| MH | 244 | 2000 |
| PU | 6531 | 2000 |
| SF | 39038 | 2000 |
|  | 11 | 2001 |
| CO | 3546 | 2001 |
| CP | 45 | 2001 |
| LH | 22 | 2001 |
| MH | 181 | 2001 |
| PU | 5470 | 2001 |
| SF | 40725 | 2001 |
|  | 3 | 2002 |
| CO | 3399 | 2002 |
| CP | 48 | 2002 |
| LH | 12 | 2002 |
| MH | 274 | 2002 |
| PU | 5053 | 2002 |
| SF | 41211 | 2002 |
|  | 14 | 2004 |
| CO | 3616 | 2004 |
| CP | 210 | 2004 |
| LH | 35 | 2004 |
| MH | 529 | 2004 |
| PU | 6829 | 2004 |
| SF | 38767 | 2004 |

* 1. No missing values for most of the years.
  2. Very few missing values observed for years 2000, 2001, 2002 and 2004.
  3. Replaced the missing values with the mode (“SF” as observed) because most number of records are categorized as Single Family Home (77% to 82%)

## POSTAL CODE:

|  |  |  |
| --- | --- | --- |
| POSTAL CODE | COUNT | YEAR |
|  | 1 | 1999 |
|  | 72 | 2000 |
|  | 1 | 2001 |
|  | 1 | 2002 |
|  | 0 | 2003 |
|  | 0 | 2004 |
|  | 1 | 2005 |
|  | 0 | 2006 |
|  | 0 | 2007 |
|  | 0 | 2008 |
|  | 0 | 2009 |
|  | 0 | 2010 |
|  | 0 | 2011 |
|  | 0 | 2012 |
|  | 0 | 2013 |
|  | 0 | 2014 |
|  | 0 | 2015 |
|  | 0 | 2016 |

* 1. 72 of 50000 unknowns in 2000, 1 row each in 1999, 2001,2002 and 2005 of unknowns
  2. Replaced the blanks with 99999 as unknown value
  3. Future Scope: Get a complete dictionary of Metropolitan Statistical Area or Metropolitan Division codes and map the MSA or MD for the row to the dictionary to find the missing postal code

## LOAN SEQUQENCE NUMBER:

* 1. Unique Identifier Column.
  2. No missing values. If the value is missing for a row, then replace by random Loan sequence number the complete row or generating a unique identifier UUID
  3. Derived two new columns for origination year and origination quarter

## LOAN PURPOSE:

* 1. No missing values in the sample files.
  2. If the values are missing then, loan purpose is unknown. Assuming that the percentage of such occurrence in the yearly data would be close (if not equal to) 0%, and it wouldn’t affect the distribution of the data, we replaced it by the mode of the column

## ORIGINAL LOAN TERM:

* 1. No missing values observed.

## NUMBER OF BORROWERS:

|  |  |  |
| --- | --- | --- |
| NUMBER OF BORROWERS | COUNT | YEAR |
|  | 30 | 1999 |
|  | 20 | 2000 |
|  | 11 | 2001 |
|  | 9 | 2002 |
|  | 7 | 2003 |
|  | 14 | 2004 |
|  | 17 | 2005 |
|  | 17 | 2006 |
|  | 23 | 2007 |
|  | 19 | 2008 |
|  | 6 | 2009 |
|  | 0 | 2010 |
|  | 0 | 2011 |
|  | 0 | 2012 |
|  | 0 | 2013 |
|  | 0 | 2014 |

* 1. 0% to 0.6% Missing values found.
  2. Replacing missing values with the mode.

## SELLER NAME:

* 1. No missing values found in the sample files.
  2. Replacing missing values by “Unknown”

## SERVICES NAME:

* 1. No missing values found in the sample files.
  2. Replacing missing values by “Unknown”

## SUPER CONFORMING FLAG:

|  |  |  |
| --- | --- | --- |
| SUPER CONFORMING FLAG | COUNT | YEAR |
| Y | 80 | 2008 |
| Y | 1236 | 2009 |
| Y | 1364 | 2010 |
| Y | 1967 | 2011 |
| Y | 2189 | 2012 |
| Y | 1718 | 2013 |
| Y | 1995 | 2014 |
| Y | 2223 | 2015 |
| Y | 492 | 2016 |

1. Per the data dictionary, all the missing values are Not super conforming, so replaced the missing values by “N”

# PERFORMANCE FILE

## LOAN SEQUENCE NUMBER:

* 1. Derived two new columns for origination year and origination quarter

## MONTHLY REPORTING PERIOD:

* 1. Derived two new columns for monthly reporting period year and month

## CURRENT ACTUAL UPB:

## CURRENT LOAN DELINQUENCY STATUS:

* 1. No Missing values observed in the sample files.
  2. Replacing missing values with “XX” which is also used for unknown.

## LOAN AGE:

* 1. No missing values observed.

## REMAINING MONTHS TO LEGAL MATURITY:

* 1. No missing values found.

## REPURCHASE FLAG:

* 1. This field is only populated at loan termination. For all others the value is not applicable.
  2. Replacing nulls with NA.

## MODIFICATION FLAG:

* 1. Replacing nulls with “NO” (Not modified)

## ZERO BALANCE CODE:

* 1. Replacing nulls and spaces with “NA” as it is not applicable if the balance is not reduced to zero.

## ZERO BALANCE EFFECTIVE DATE:

* 1. Replacing missing values with 999999, which will denote not applicable.
  2. Deriving 2 new columns for zero balance effective year and month.

## CURRENT INTEREST RATE:

* 1. Replacing empty values with 0.

## DUE DATE OF LAST PAID INSTALLMENT:

* 1. Replacing missing values with 999999.
  2. Deriving 2 new columns for due year and month of last paid installment.

## Replacing missing values with 0 for the following columns

* 1. MI RECOVERIES
  2. NET SALES PROCEEDS
  3. NON MI RECOVERIES
  4. EXPENSES
  5. LEGAL COSTS
  6. MAINTENANCE AND PRESERVATION COSTS:
  7. TAXES AND INSURANCE:
  8. MISCELLENEOUS EXPENSES:
  9. ACTUAL LOSS CALCULATION:
  10. MODIFICATION COST
  11. CURRENT DEFERRED UPB

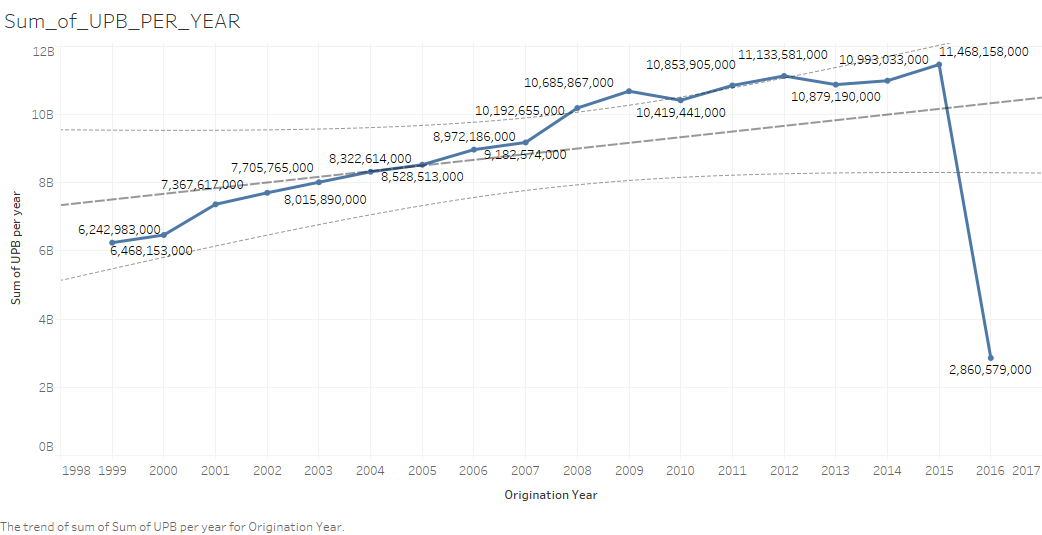
# Exploratory Data analysis

### SUMMARIES OF SAMPLE ORIGINAL FILES- 1999

## Sum of Original UPB PER YEAR (1999)

ORIGINATION YEAR Sum\_of\_UPB\_per\_year

0 99 6242983000



# OBSERVATION

We see an inclination for the sum of UPB each year, which is natural considering the inflation and demand each year. We also observe a decline during 2009, which would be the effect of the financial crisis of ‘08 and ‘09.

## Sum of Original UPB per QUARTER

ORIGINATION QUARTER Sum of UPB per Quarter

0 1 1577407000

1 2 1584374000

2 3 1541836000

3 4 1539366000

## Sum of Original UPB per YEAR PER Quarter

ORIGINATION YEAR ORIGINATION QUARTER Sum of UPB /year /quarter

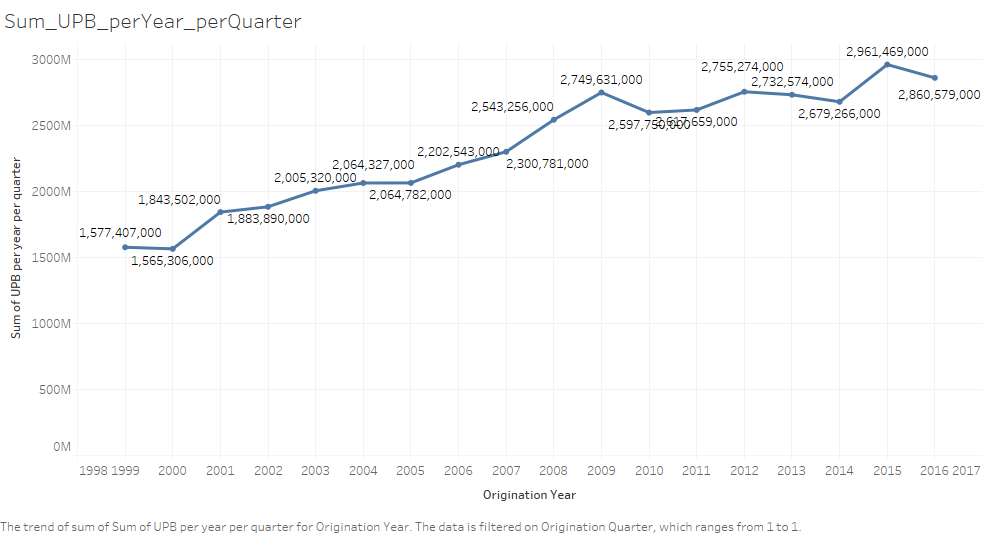
0 99 1 1577407000

1 99 2 1584374000

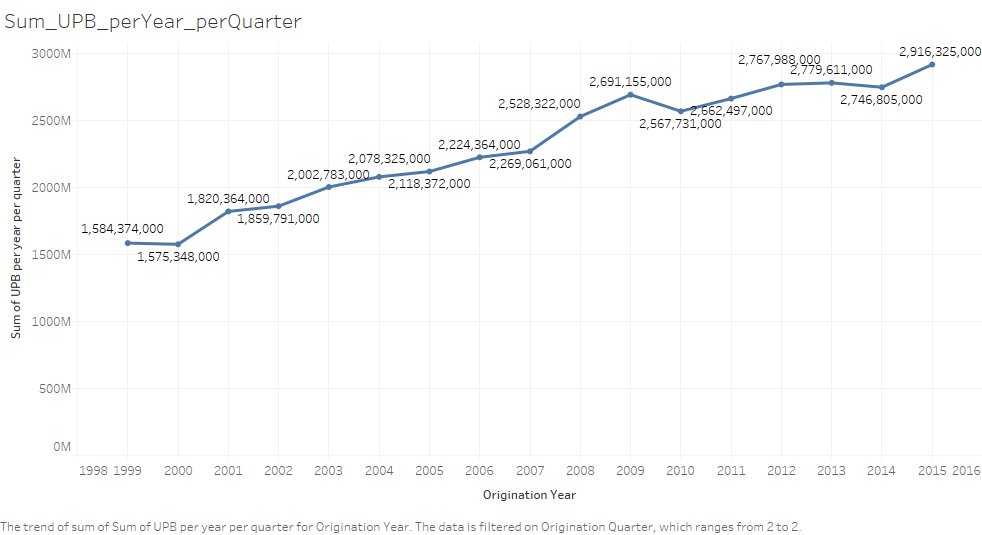
2 99 3 1541836000

3 99 4 1539366000

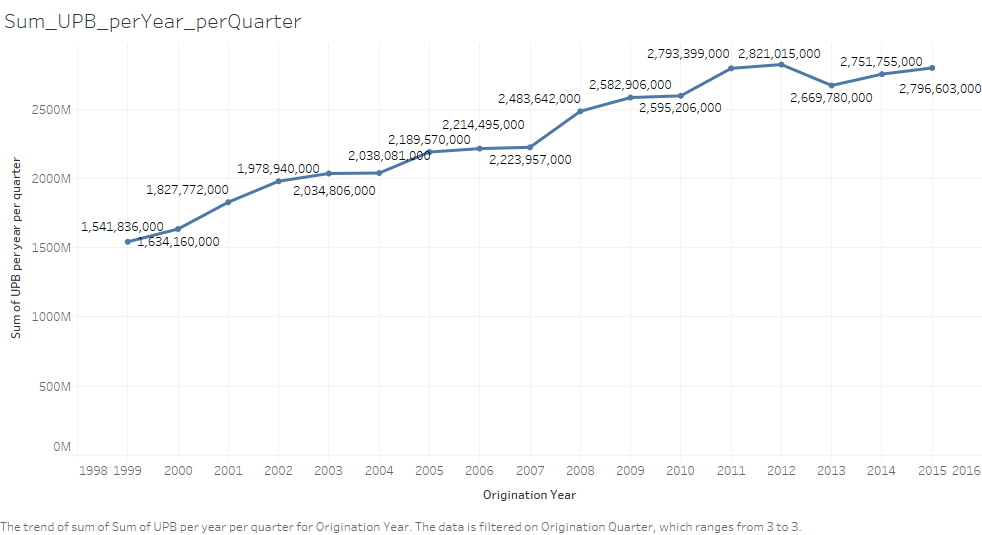
## Quarter 1



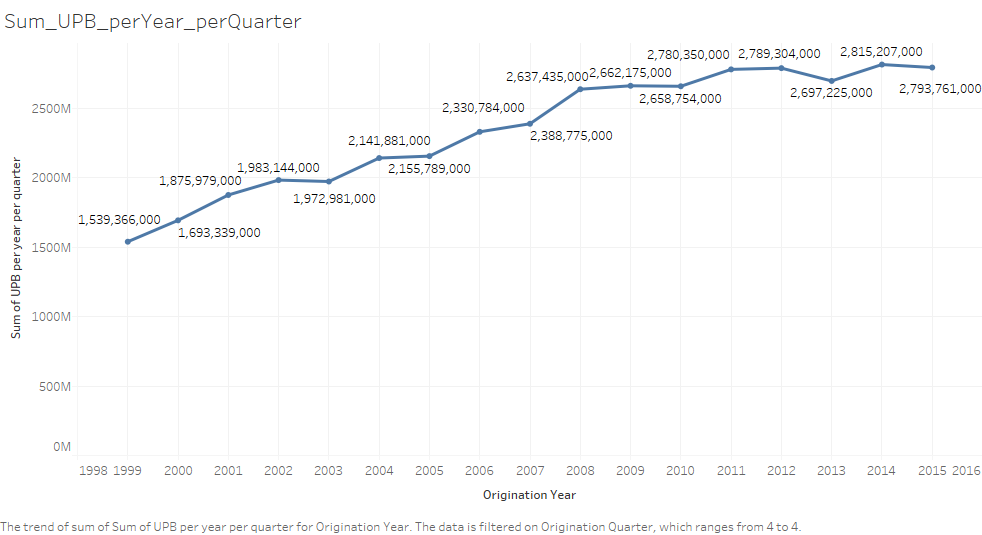
## Quarter 2



## Quarter 3



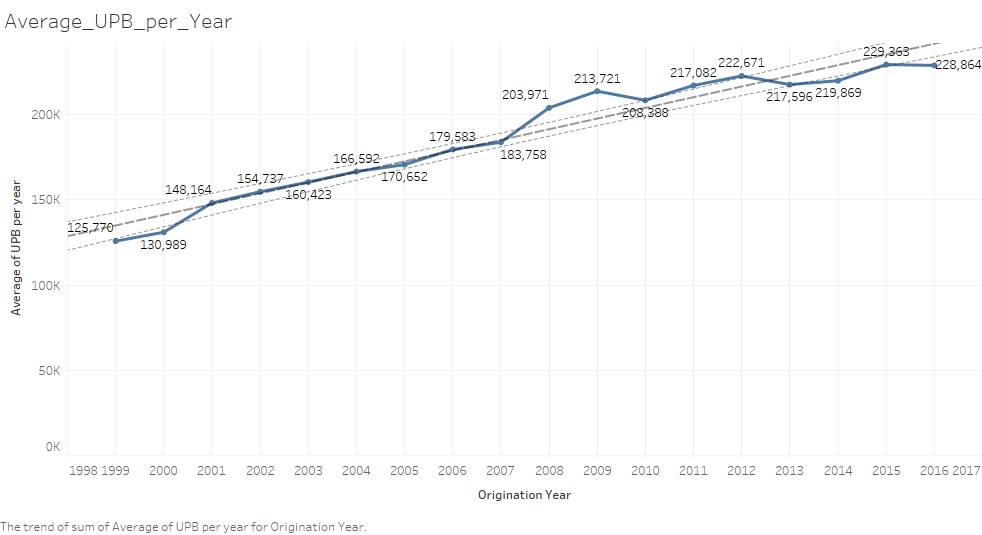
## Quarter 4



## AVERAGE of Original UPB per year

ORIGINATION YEAR Average of UPB per year

0 99 125770



Similar observation as the sum of UPB. Inclination in initial years and a declination in 2009

## Average of Original UPB per year per Quarter

ORIGINATION QUARTER Average of UPB per quarter

0 1 126699

1 2 127607

2 3 124492

3 4 124272

## Average of Original UPB per year per Quarter

ORIGINATION YEAR ORIGINATION QUARTER Average of UPB per year per quarter

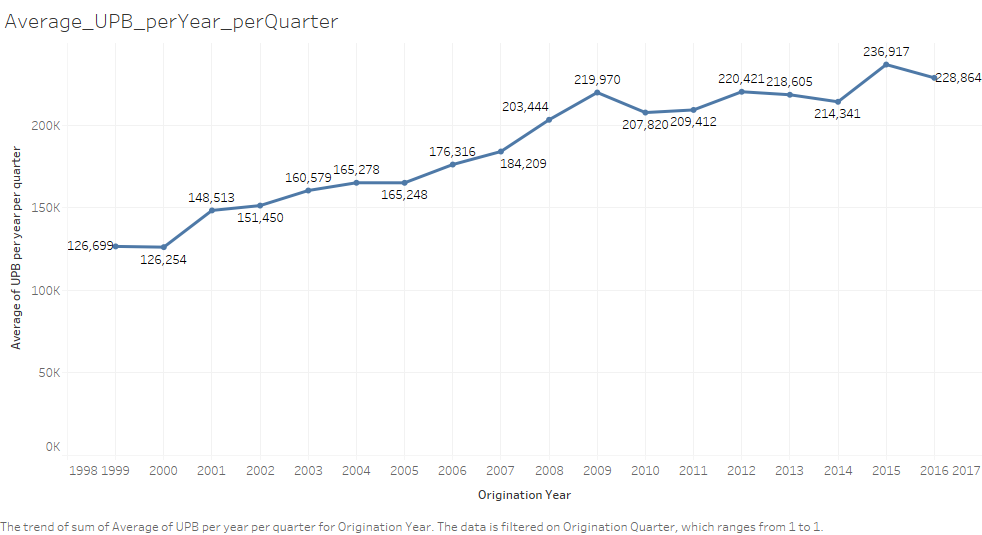
0 99 1 126699

1 99 2 127607

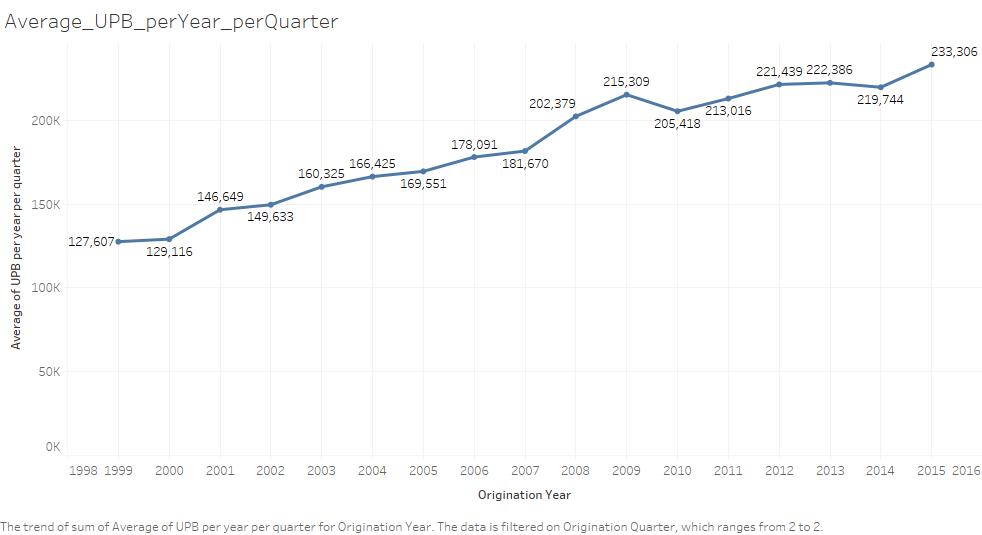
2 99 3 124492

3 99 4 124272

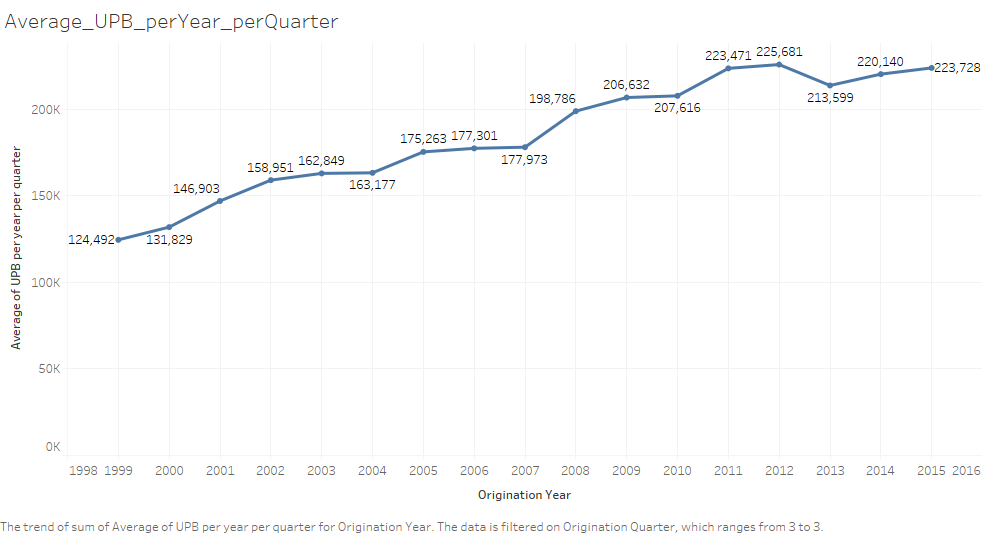
## QUARTER 1



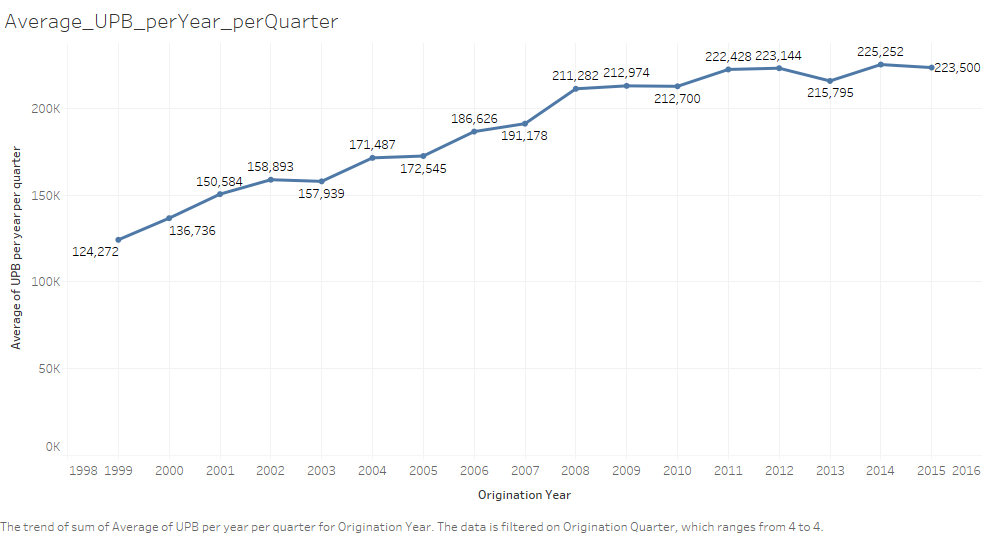
## QUARTER 2



## QUARTER 3



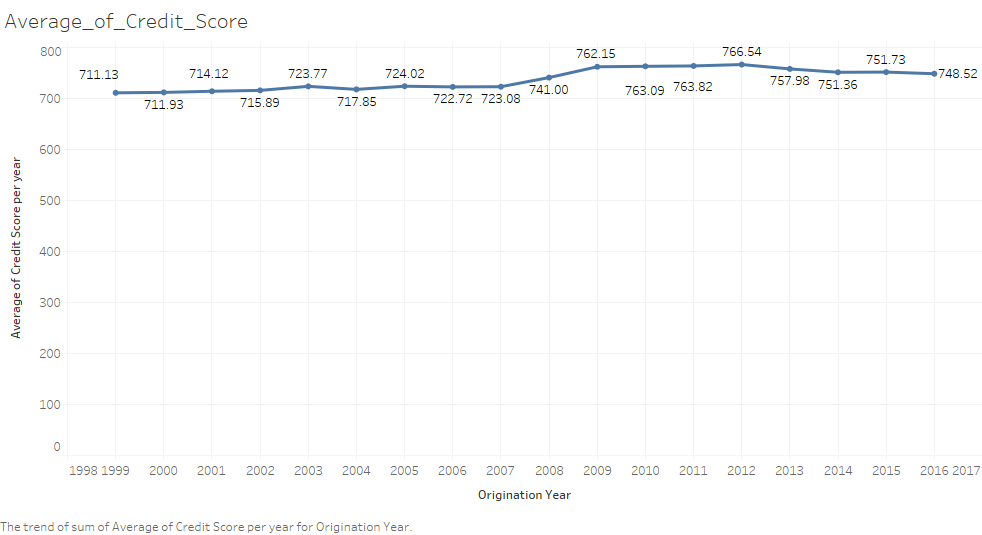
## QUARTER 4



## Average of Credit Score per year

ORIGINATION YEAR Average of Credit Score per year

0 99 711.133748



# OBSERVATION

Initially we observe that the loan is given to people with avg. credit score of 711-741, but after financial crisis of ’08 and ’09 we observe a significant increase in the credit score, which means that the banks were not willing to take high risk

## Average of Credit Score per Quarter

ORIGINATION QUARTER Average\_Credit\_Score\_per\_quarter

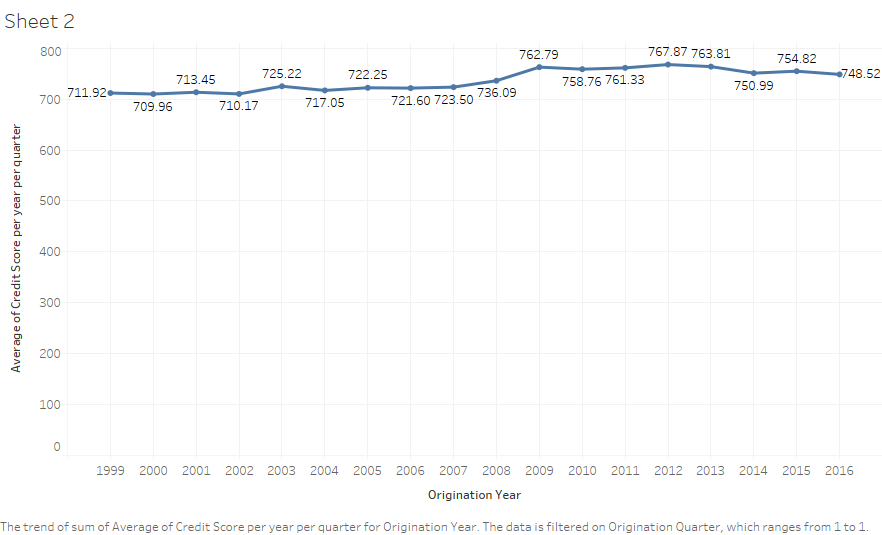
0 1 711.919036

1 2 711.671472

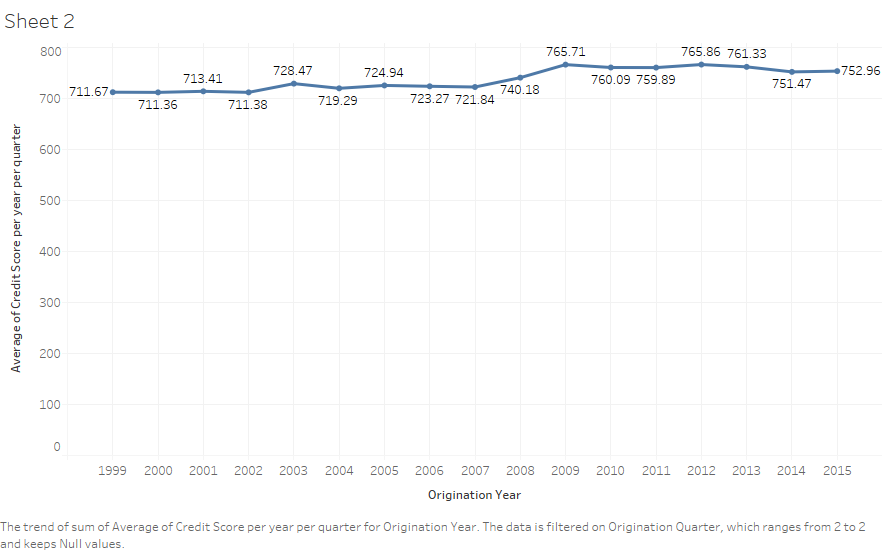
2 3 711.183044

3 4 709.756196

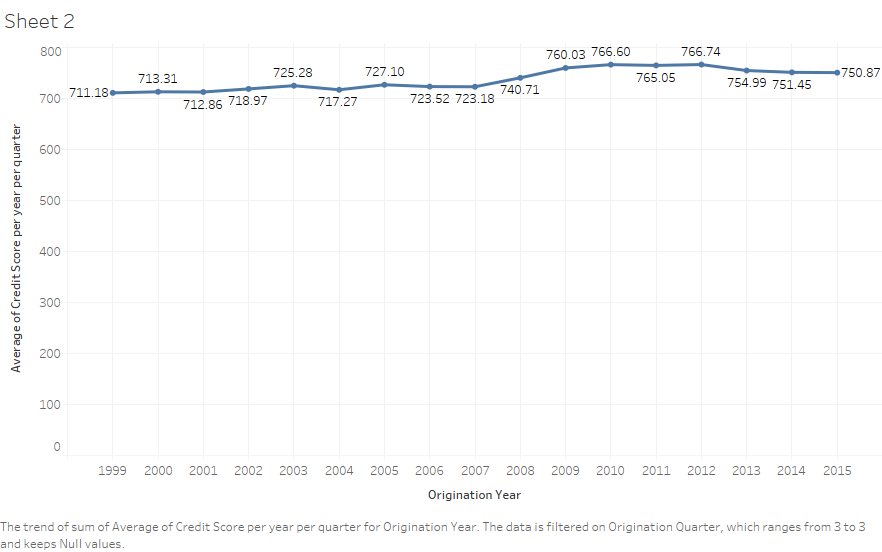
## QUARTER 1



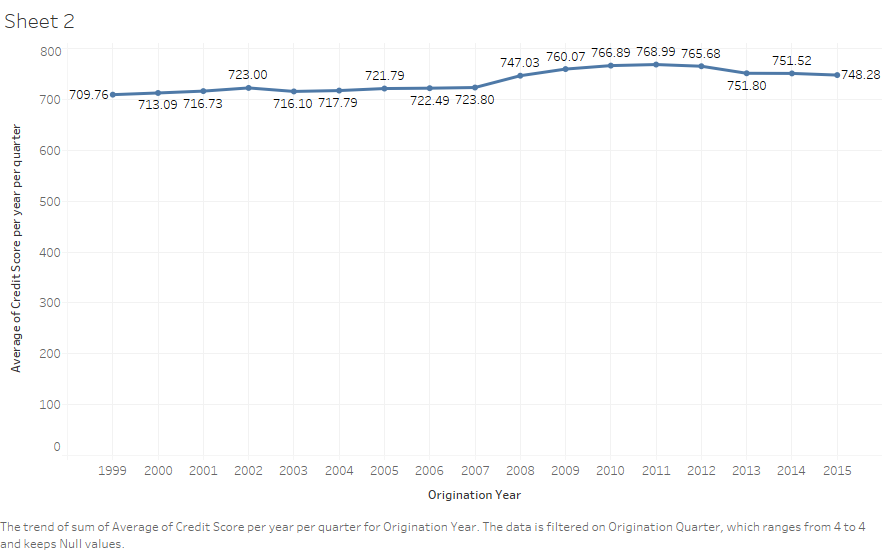
## QUARTER 2



## QUARTER 3



## QUARTER 4



## Average of Credit Score per year per Quarter

ORIGINATION\_YEAR ORIGINATION\_QUARTER Average\_Credit\_Score\_year\_quarter

0 99 1 711.919036

1 99 2 711.671472

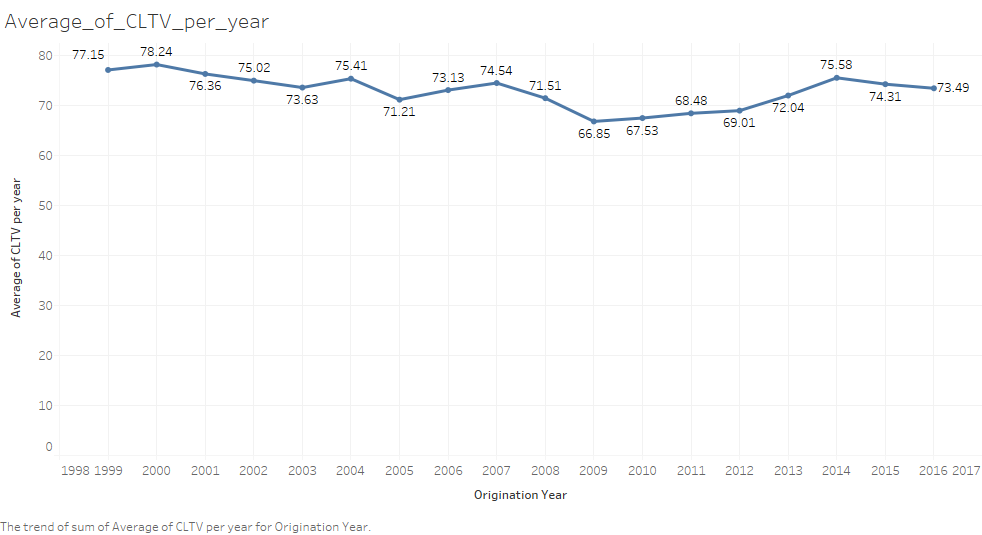
2 99 3 711.183044

3 99 4 709.756196

## Average of CLTV per year

ORIGINATION\_YEAR Average of CLTV per year

0 99 77.153753



## Average of CLTV per Quarter

ORIGINATION\_QUARTER Average\_CLTV\_quarter

0 1 75.212610

1 2 77.176627

2 3 77.813726

3 4 78.421975

ORIGINATION\_YEAR ORIGINATION\_QUARTER

99 1

99 2

99 3

99 4

## Average of CLTV per year per Quarter

Average of CLTV Score per year per quarter

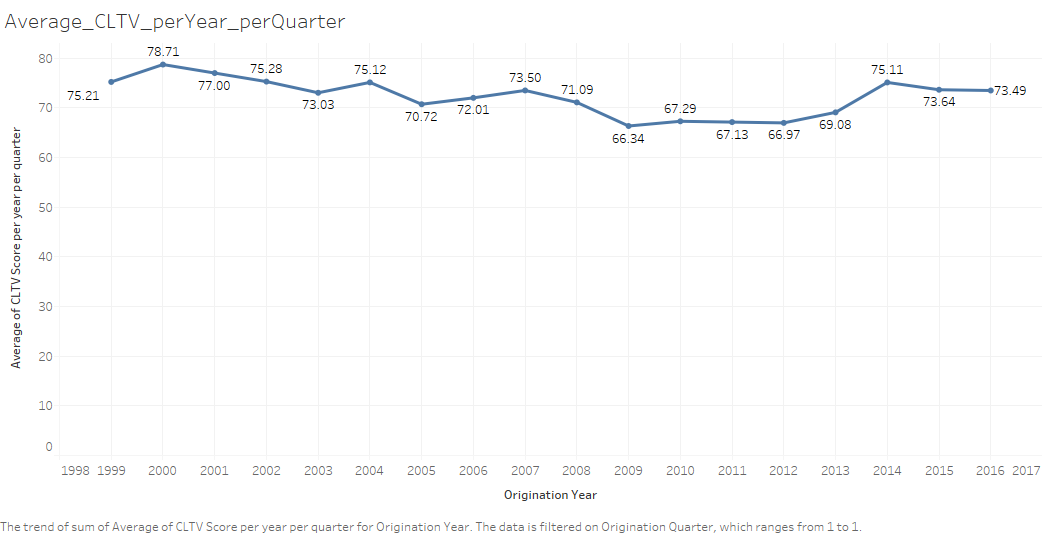
0 75.212610

1 77.176627

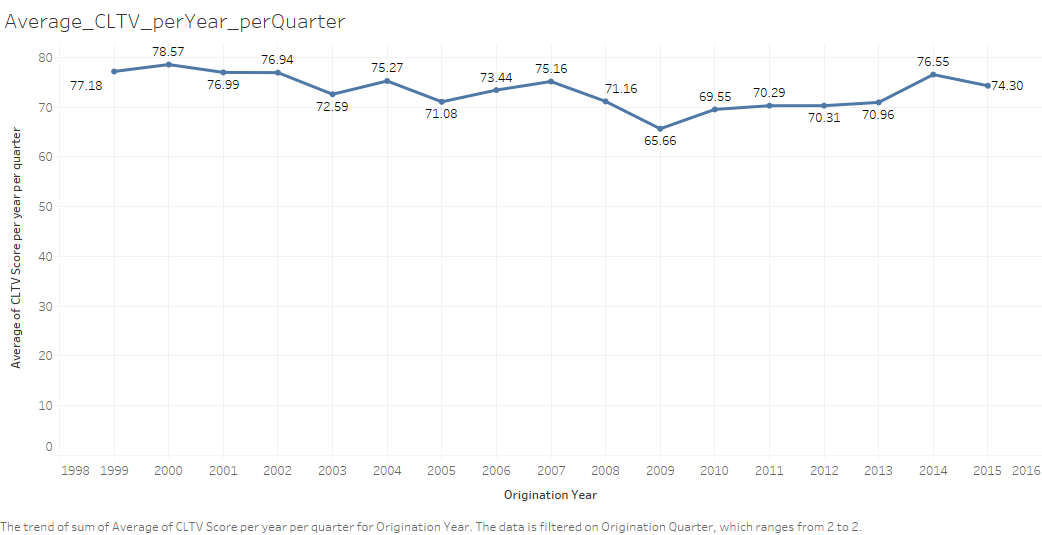
2 77.813726

3 78.421975

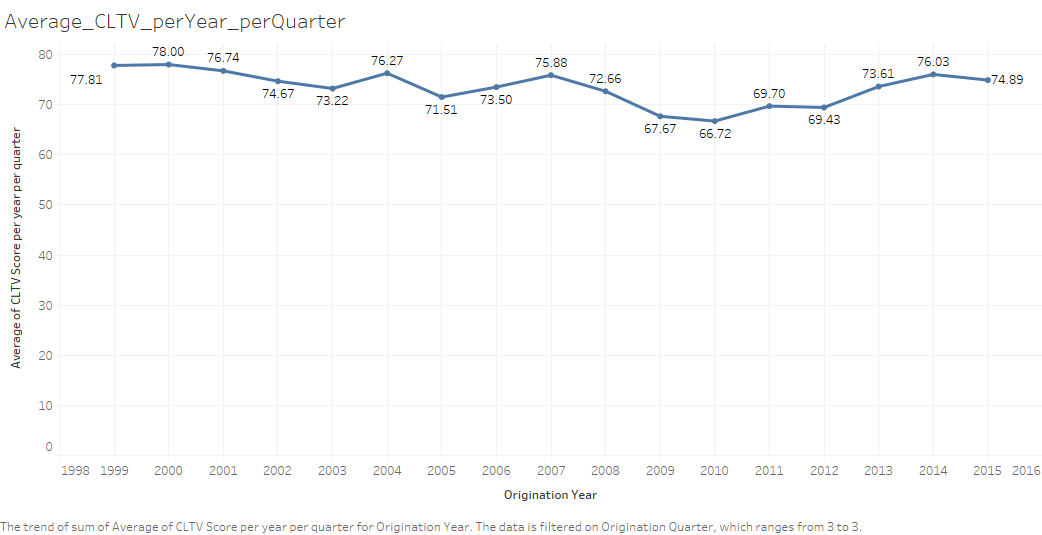
## QUARTER 1



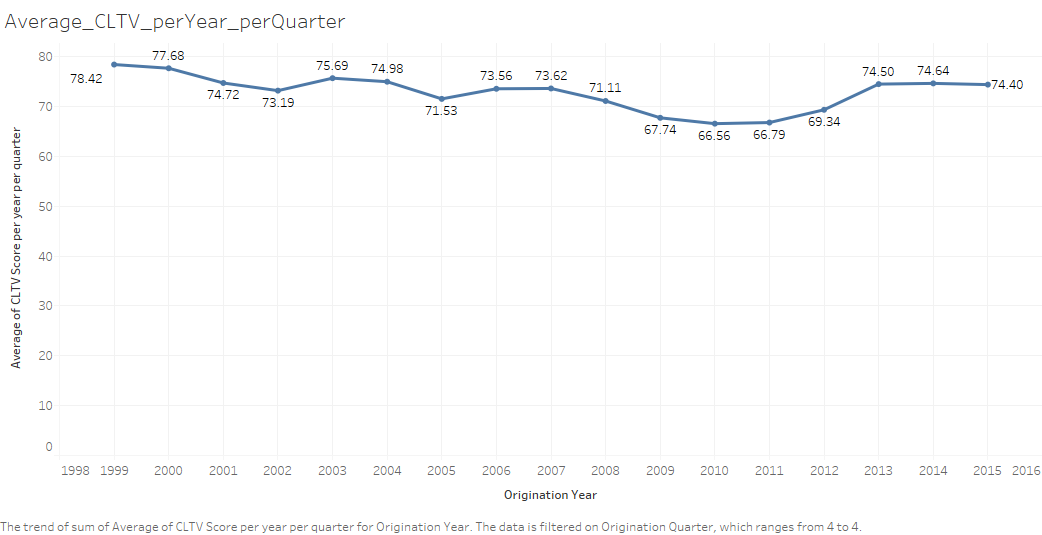
## QUARTER 2



## QUARTER 3



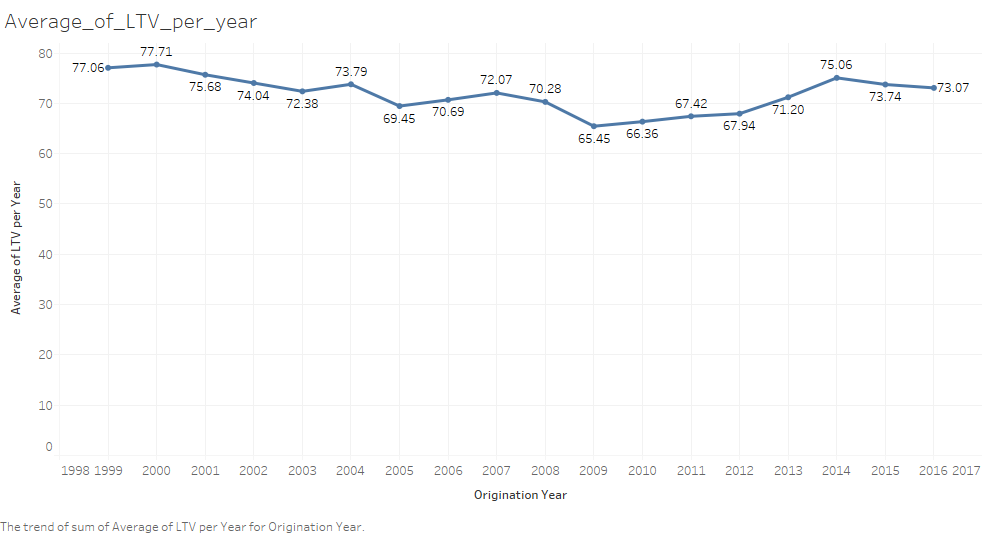
## QUARTER 4



## Average of LTV per Year

ORIGINATION\_YEAR Average\_LTV\_per\_Year

0 99 77.056509



## Average of LTV per Quarter

ORIGINATION\_QUARTER Average\_LTV\_Quarter

0 1 75.155020

1 2 77.108811

2 3 77.715382

3 4 78.256479

## Average of LTV per year per Quarter

ORIGINATION\_YEAR ORIGINATION\_QUARTER

0 99 1

1 99 2

2 99 3

3 99 4

Average\_LTV\_Score\_year\_quarter

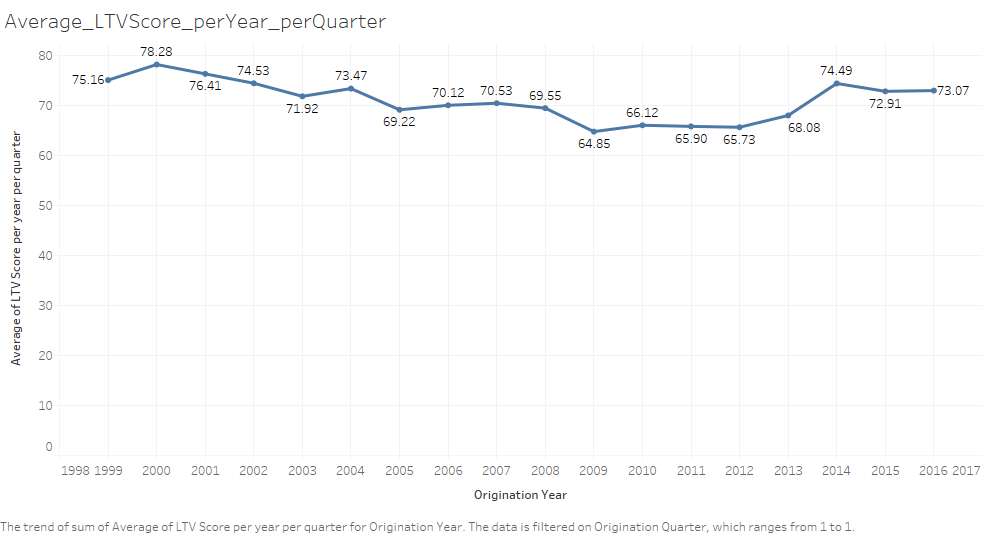
75.155020

77.108811

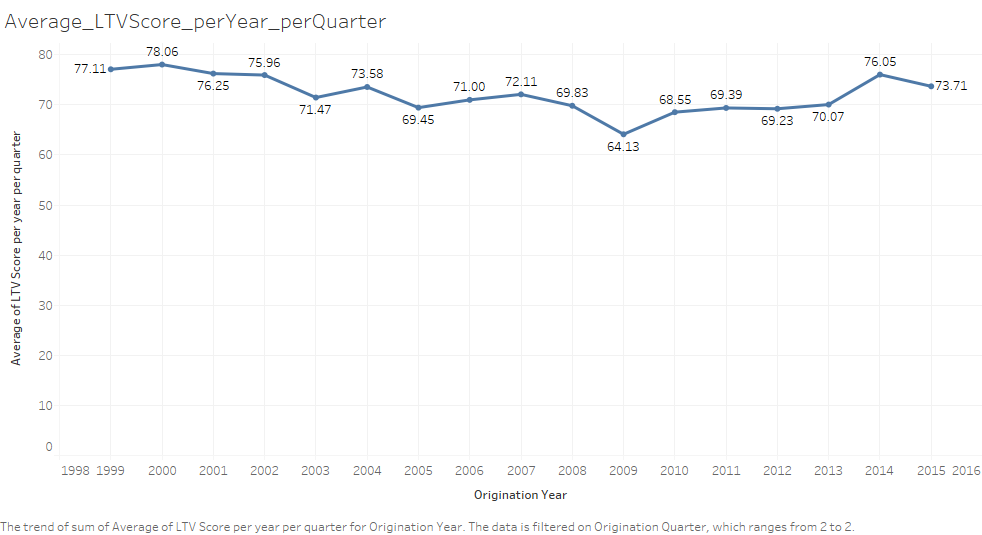
77.715382

78.256479

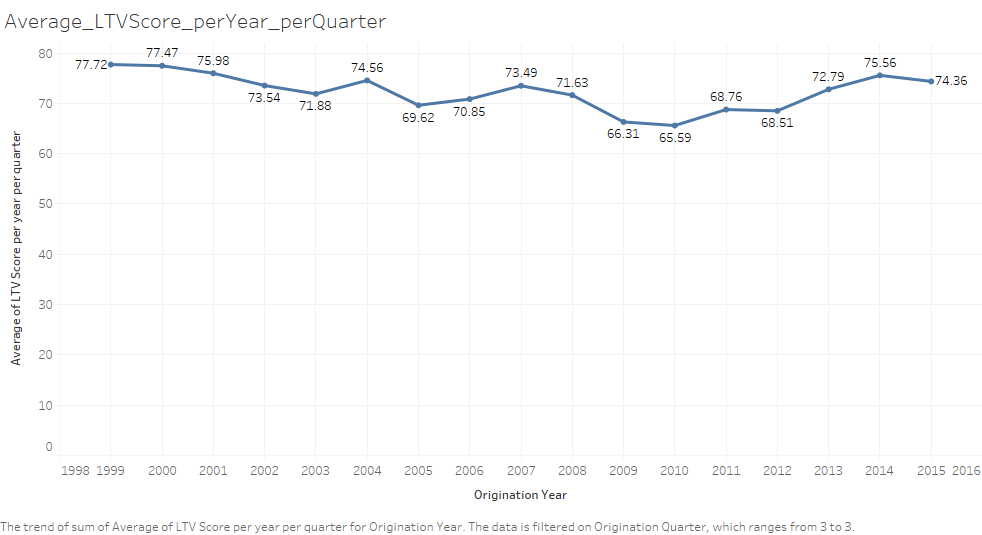
## QUARTER 1



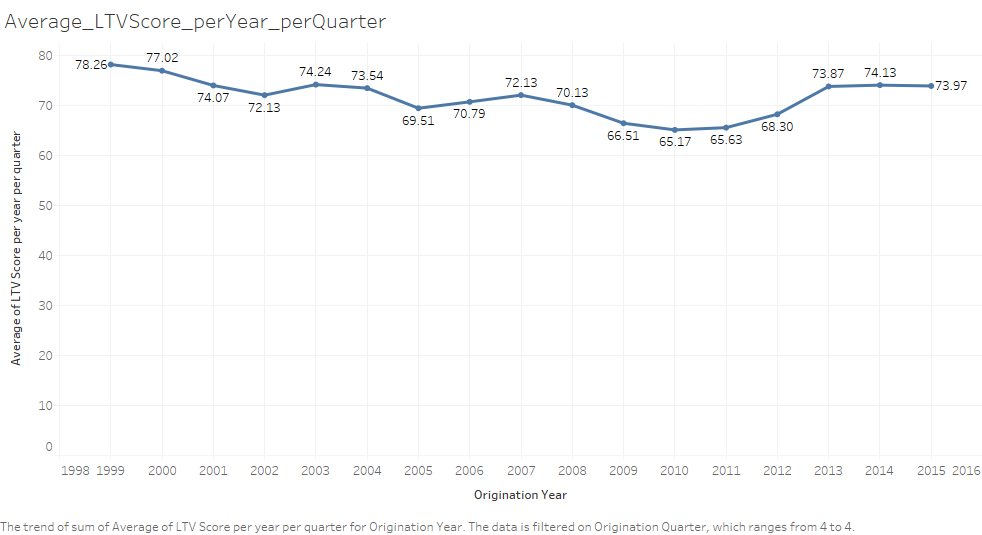
## QUARTER 2



## QUARTER 3



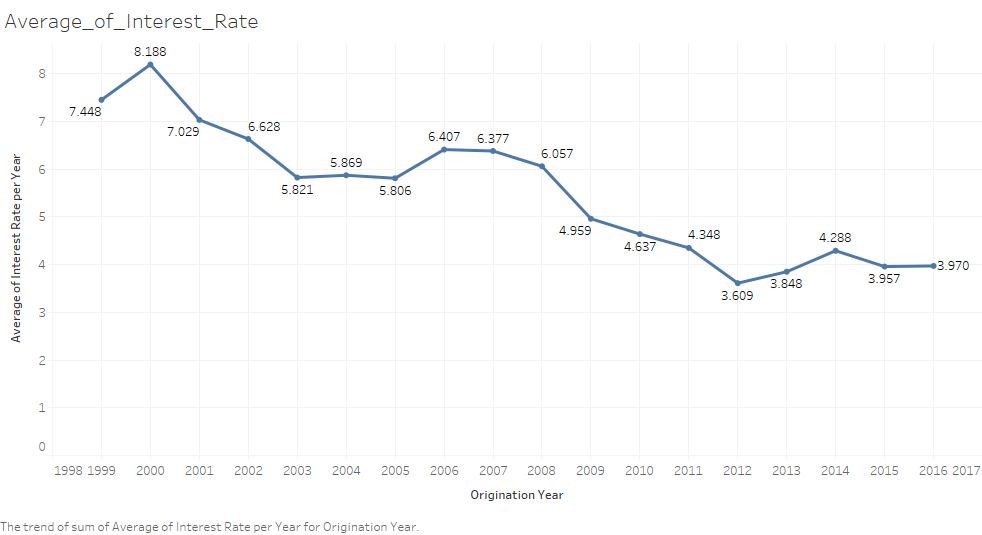
## QUARTER 4



## Average of Interest Rate per Year

ORIGINATION\_YEAR Average\_Interest\_Rate\_per\_Year

0 99 7.447759



# OBSERVATION

We see that the interest rate is declining over the years. After looking at the statistics of the financial years, we can say that the interest rate is the higher when the economy is good, and declines when bad. Hence, 2012 has the least and improves during 2013.

## Average of Interest Rate per Quarter

ORIGINATION\_QUARTER Average\_Interest\_Rate\_per\_Quarter

0 1 6.927040

1 2 7.155169

2 3 7.777791

3 4 7.934422

## Average of Interest Rate per year per Quarter

ORIGINATION\_YEAR ORIGINATION\_QUARTER Average\_Interest\_rate\_year\_quarter

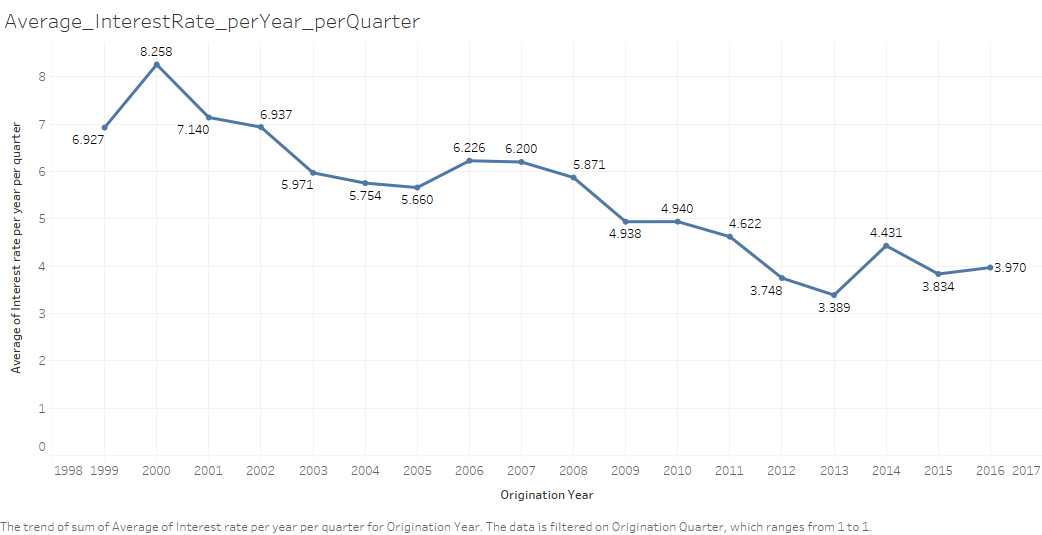
0 99 1 6.927040

1 99 2 7.155169

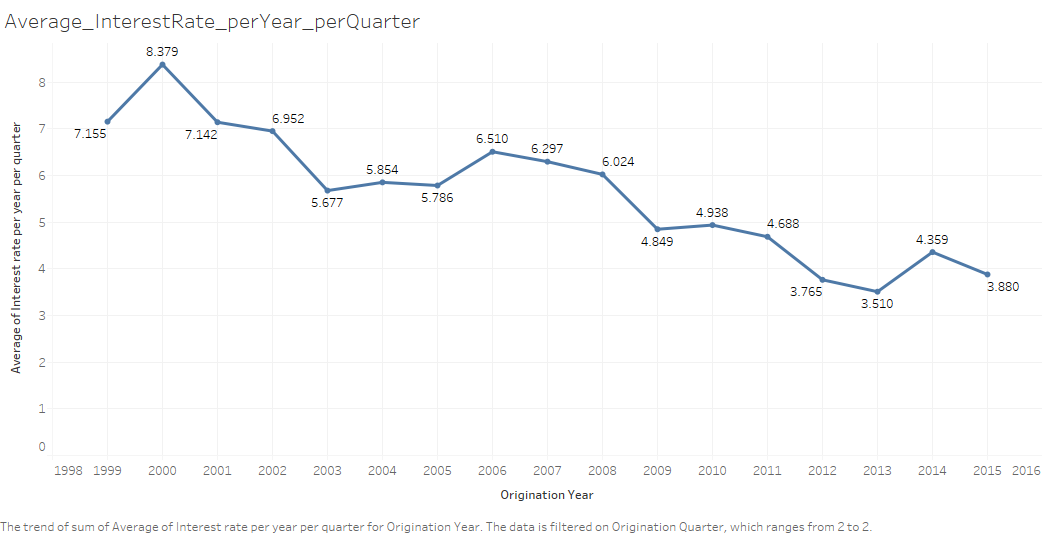
2 99 3 7.777791

3 99 4 7.934422

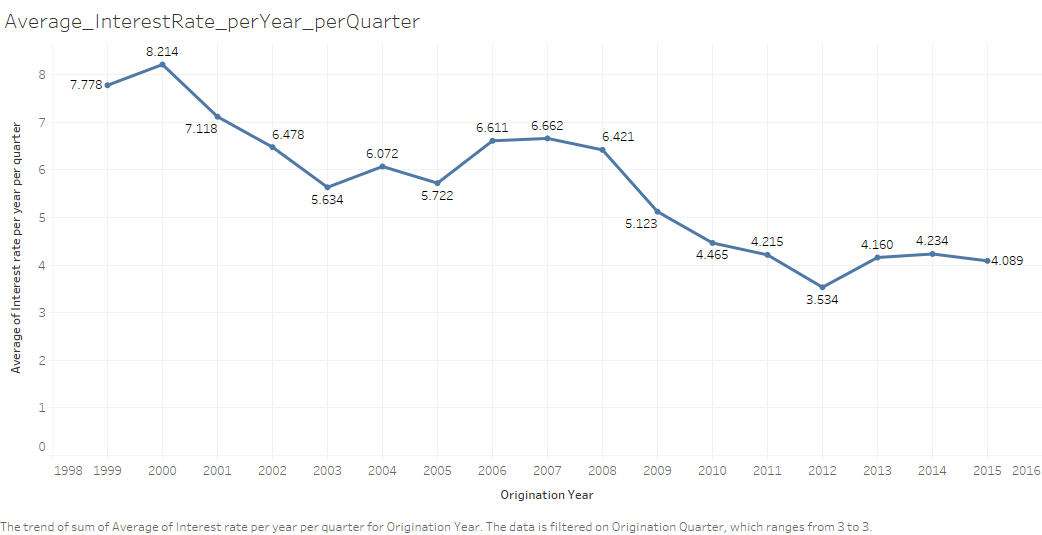
## QUARTER 1



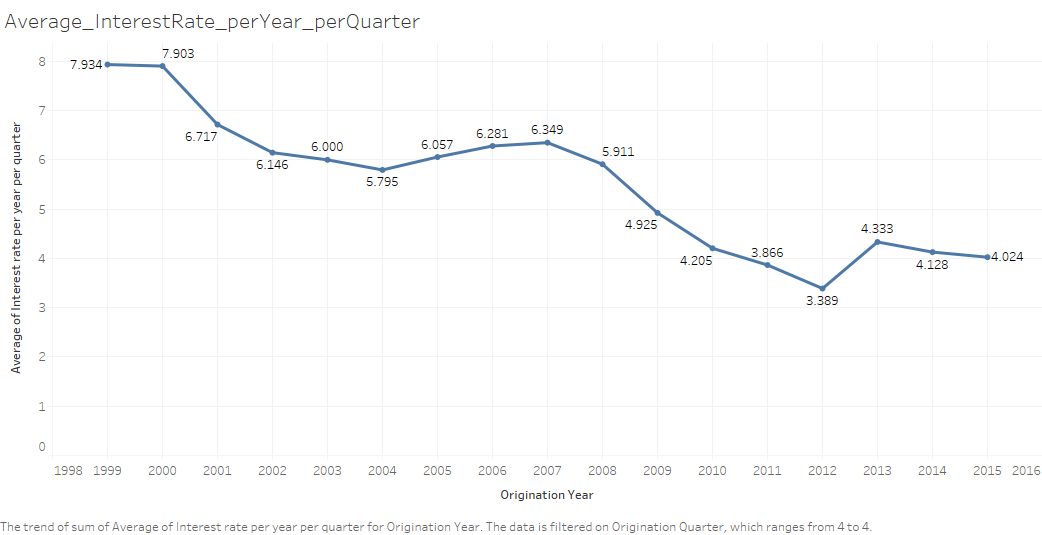
## QUARTER 2



## QUARTER 3

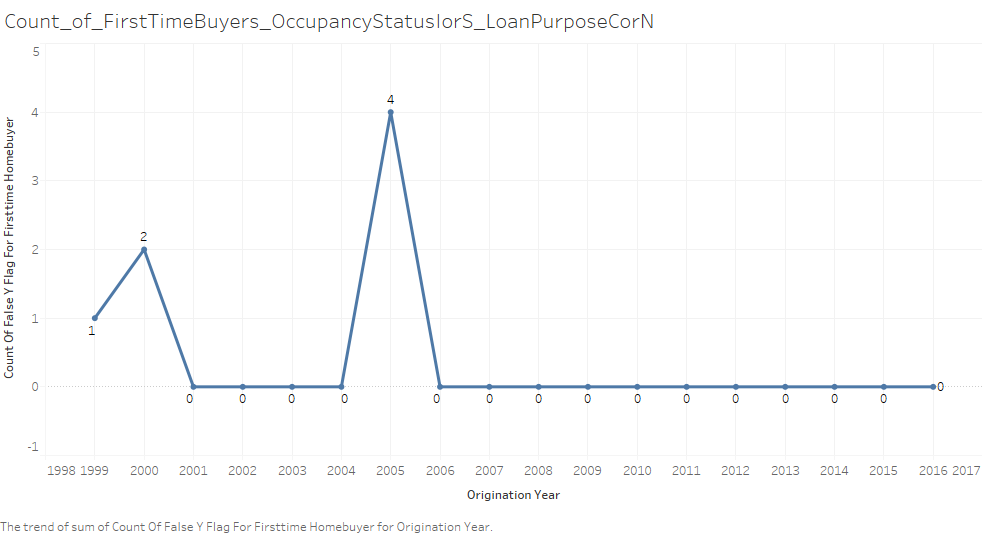


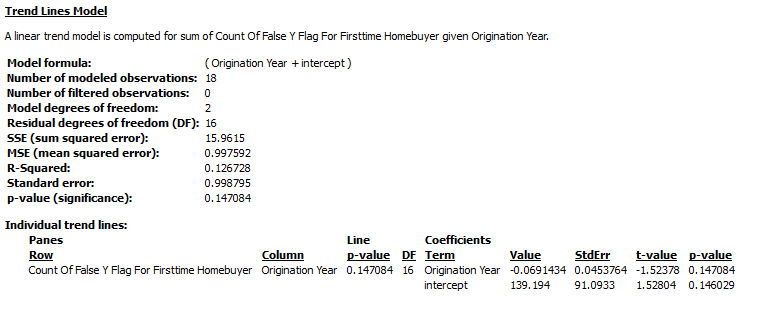
## QUARTER 4



## Count of Loans with First Time Home Buyer equal to "Y", Occupancy equal to "I" or "S" and Loan Purpose equal to "C" and "N"

1





## Count of Loans with MSA flag equal to "YES"

42069

## Count of Loans with MSA flag equal to "NO"

7569

## Average Original UPB where MSA flag equal to "YES"

129213.17359575935

## Average Original UPB where MSA flag equal to "YES"

106634.1656757828

## Average Credit Score where MSA flag equal to "YES"

711.3050464712734

## Average Credit Score where MSA flag equal to "NO"

710.1816620425419

## Average Interest Rate where MSA flag equal to "YES"

7.434357864460766

## Average Interest Rate where MSA flag equal to "NO"

7.522243361078074

## Count of Loans where PPM flag equal to "Y"

262

# SUMMARIES OF SAMPLE PERFORMANCE FILES- 2016

## DISTINCT COUNT OF LOAN

LOAN SEQUENCE NUMBER Distinct count of loan

0 F116Q1000017 8

1 F116Q1000025 5

2 F116Q1000034 8

3 F116Q1000051 8

4 F116Q1000076 8

5 F116Q1000093 8

6 F116Q1000110 7

7 F116Q1000186 8

8 F116Q1000203 8

9 F116Q1000212 8

10 F116Q1000220 7

11 F116Q1000229 8

12 F116Q1000245 8

13 F116Q1000262 8

14 F116Q1000271 8

15 F116Q1000288 8

16 F116Q1000296 7

17 F116Q1000330 8

18 F116Q1000381 8

19 F116Q1000389 8

20 F116Q1000423 8

21 F116Q1000440 5

22 F116Q1000508 8

23 F116Q1000525 8

24 F116Q1000559 8

25 F116Q1000584 8

26 F116Q1000609 8

27 F116Q1000660 8

28 F116Q1000669 8

29 F116Q1000686 8

... ... ...

12444 F116Q1276390 3

12445 F116Q1276415 3

12446 F116Q1276424 3

12447 F116Q1276441 3

12448 F116Q1276449 3

12449 F116Q1276458 3

12450 F116Q1276475 3

12451 F116Q1276492 3

12452 F116Q1276500 3

12453 F116Q1276509 3

12454 F116Q1276517 4

12455 F116Q1276526 4

12456 F116Q1276542 4

12457 F116Q1276551 4

12458 F116Q1276559 4

12459 F116Q1276568 4

12460 F116Q1276585 3

12461 F116Q1276602 3

12462 F116Q1276610 3

12463 F116Q1276619 3

12464 F116Q1276627 3

12465 F116Q1276644 3

12466 F116Q1276652 3

12467 F116Q1276669 3

12468 F116Q1276686 3

12469 F116Q1276695 2

12470 F116Q1276729 2

12471 F116Q1276779 1

12472 F116Q1276813 1

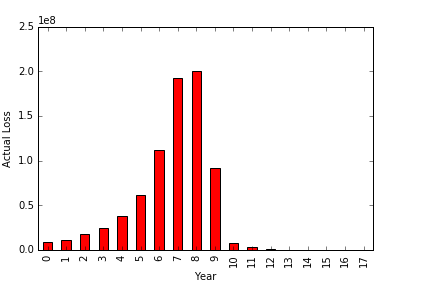
12473 F116Q1276864 1

## COUNT OF LOANS WITH CURRENT UPB EQUAL TO 0, AND ZERO BALANCE CODE EQUAL TO 1 OR 6

424

## Count of ACTUAL LOSS CALCULATION with Current UPB equal to 0, and Zero balance code equal to 9

0



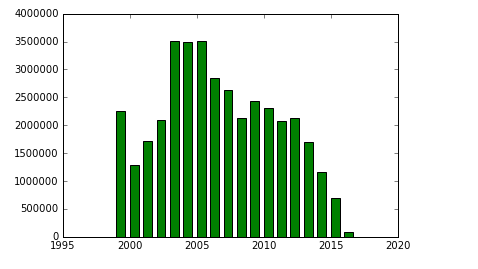
# 0BSERVATION

Here, we are calculating the actual losses for the loans that have been closed due to non-performance, and sold or auctioned by the lender.

We see that there is a significant increase from 2006 through 2008

## COUNT OF Loans with Current UPB not equal to 0

83058



# 0BSERVATION

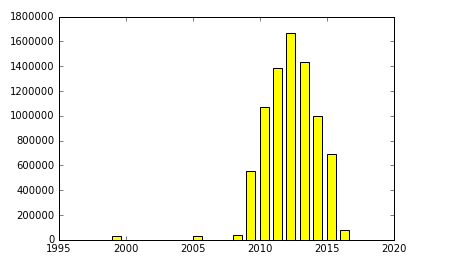
These are the current active loans

## Count of Loans with DELINQUENCY STATUS >= 5

0

## Count of Loans with Current UPB not equal to 0 and DELINQUENCY STATUS >= 0

82917



# 0BSERVATION

With this observation, we can infer that the loans are active and has been delinquent at least once

## Count of Modification flag grouped by Loan Sequence Number

LOAN SEQUENCE NUMBER \

0 F116Q1000017

1 F116Q1000025

2 F116Q1000034

3 F116Q1000051

4 F116Q1000076

5 F116Q1000093

6 F116Q1000110

7 F116Q1000186

8 F116Q1000203

9 F116Q1000212

10 F116Q1000220

11 F116Q1000229

12 F116Q1000245

13 F116Q1000262

14 F116Q1000271

15 F116Q1000288

16 F116Q1000296

17 F116Q1000330

18 F116Q1000381

19 F116Q1000389

20 F116Q1000423

21 F116Q1000440

22 F116Q1000508

23 F116Q1000525

24 F116Q1000559

25 F116Q1000584

26 F116Q1000609

27 F116Q1000660

28 F116Q1000669

29 F116Q1000686

... ...

12444 F116Q1276390

12445 F116Q1276415

12446 F116Q1276424

12447 F116Q1276441

12448 F116Q1276449

12449 F116Q1276458

12450 F116Q1276475

12451 F116Q1276492

12452 F116Q1276500

12453 F116Q1276509

12454 F116Q1276517

12455 F116Q1276526

12456 F116Q1276542

12457 F116Q1276551

12458 F116Q1276559

12459 F116Q1276568

12460 F116Q1276585

12461 F116Q1276602

12462 F116Q1276610

12463 F116Q1276619

12464 F116Q1276627

12465 F116Q1276644

12466 F116Q1276652

12467 F116Q1276669

12468 F116Q1276686

12469 F116Q1276695

12470 F116Q1276729

12471 F116Q1276779

12472 F116Q1276813

12473 F116Q1276864

Number of Modification flag grouped by Loan Sequence Number

0 8

1 5

2 8

3 8

4 8

5 8

6 7

7 8

8 8

9 8

10 7

11 8

12 8

13 8

14 8

15 8

16 7

17 8

18 8

19 8

20 8

21 5

22 8

23 8

24 8

25 8

26 8

27 8

28 8

29 8

... ...

12444 3

12445 3

12446 3

12447 3

12448 3

12449 3

12450 3

12451 3

12452 3

12453 3

12454 4

12455 4

12456 4

12457 4

12458 4

12459 4

12460 3

12461 3

12462 3

12463 3

12464 3

12465 3

12466 3

12467 3

12468 3

12469 2

12470 2

12471 1

12472 1

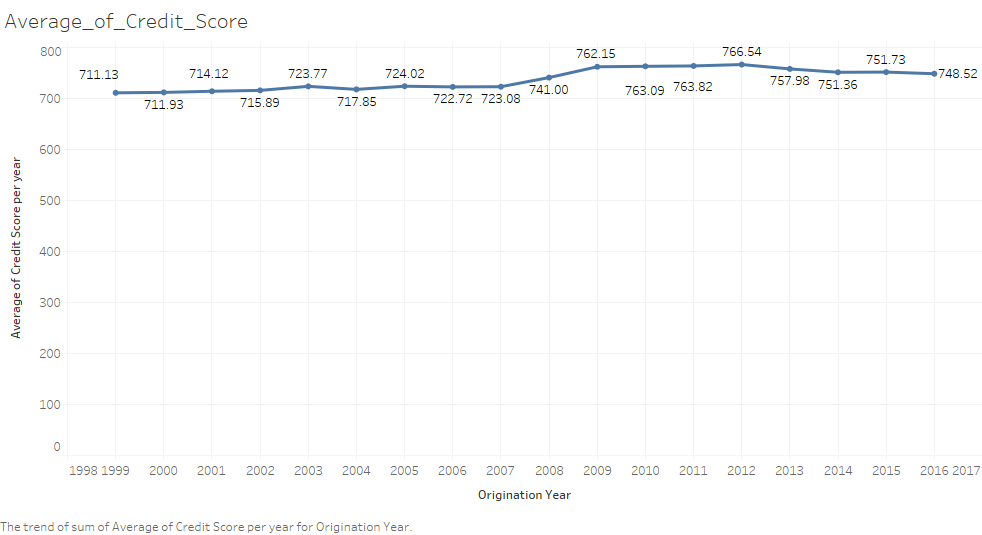
12473 1

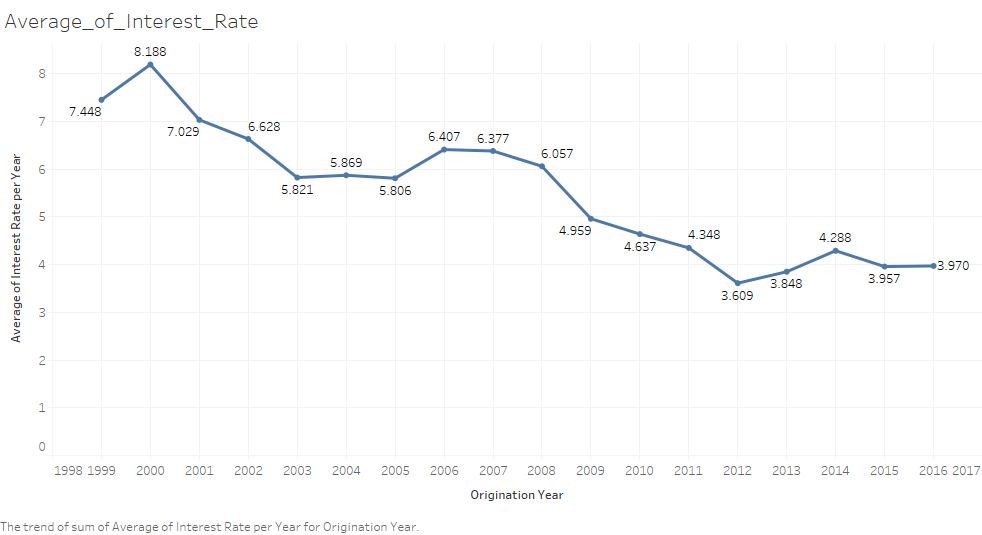
[12474 rows x 2 columns]

# **OBSERVATIONS**

1. The Trend Lines for Average Credit Score Shows that 2012 has the highest Credit Score (766.54). Meanwhile, the Trend Lines for The Average Interest Rate is the least in 2012 (3.609)

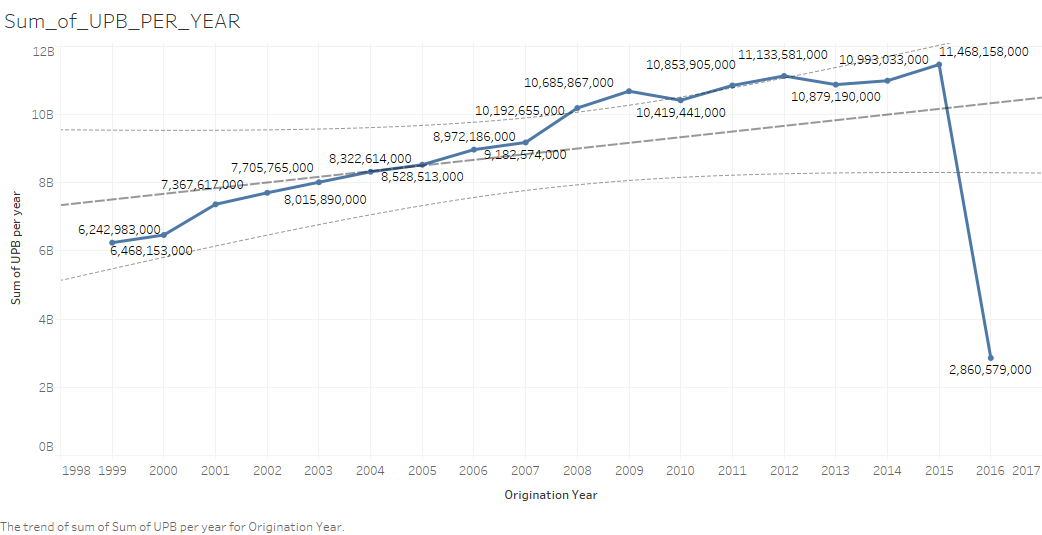
Hence, it can be inferred that the Credit Score is the highest in 2012, since the Interest rate is the least





1. We observed that the Sums and Average values for the year 2016 is much lesser than the rest of the years, since only the data for the first quarter is available

For Example: The Sum of UPB for 2016 is much lesser than the rest



**Programming Language used : Python**

**Workflow Manager User: Luigi**

**Tasks**

1. **Downloading Data**
2. **Preprocessing Origination Data**
3. **Preprocessing Performance Data**
4. **Building prediction model**

# Downloading Data(Part 2):

File Location : Classes/Part1/Download\_sf\_loan.py

Task Requires no prior tasks to be completed.

Output of the task are all the sample origination and performance files.

Process:

* Asking user for username and password.
* Creating a browser agent (using the mechanicalsoup library) to store and pass the cookies
* Logging in with the user’s credentials.
* Checking if the user is succesfuly logged in or not.
* Landing to the page that contains the list of files and download links
* Asking user for the year and the quarter file to run prediction model.
* Putting the table of files in a dataframe
* Iterating through the rows in dataframe for the links that contain sample files and downloading them to a newly created (if it doesn’t already exist) “Downloads” directory
* The program also checks if the files are already present in the “Downloads” directory. It skips the downloading if the file already exists.
* Unzipping the downloaded file.

# Origination File Observation and Cleaning

1. Credit Score: Deleted the rows that had missing credit score
   1. Cannot replace missing values as it is explicitly specified that credit score can be either less than 301 or greater than 850.
   2. Number of such instances is very less (0.002% in 2016 to 1.242% in 2000)
   3. Removing the rows that have blank values and nulls for credit score.

|  |  |  |
| --- | --- | --- |
| CREDIT SCORE | COUNT OF BLANKS | YEAR |
|  | 362 | 1999 |
|  | 621 | 2000 |
|  | 274 | 2001 |
|  | 201 | 2002 |
|  | 33 | 2003 |
|  | 42 | 2004 |
|  | 24 | 2005 |
|  | 39 | 2006 |
|  | 29 | 2007 |
|  | 29 | 2008 |
|  | 1 | 2009 |
|  | 1 | 2011 |
|  | 3 | 2013 |
|  | 2 | 2014 |
|  | 1 | 2016 |

## FIRST PAYMENT DATE:

No missing values in sample files

## FIRST TIME HOMEBUYER FLAG:

* 1. If blank it can be replaced by NA if Occupancy Status is either “I” or “S” (Investment property or Second Home)
  2. If blank it can replaced by NA if Loan Purpose is either “C” or “N” (Refinance)
  3. If blank, then replace it with NA
  4. Created three columns for – First Time Homebuyer Flag YES (1,0) , NO(1,0) and NA(1,0)

## MATURITY DATE:

* 1. No missing values in the sample files
  2. Splitting Maturity year and month

## METROPOLITAN STATISTICAL AREA(MSA) OR METROPOLITAN DIVISION:

* 1. Replaced missing values with zero.
  2. Derived a new column for Metropolitan Area Flag, that had values in it
  3. Future Scope: Compare the values of zip codes, if the zip code belongs to a MSA or MD, then map the msa or md code in the data.

|  |  |
| --- | --- |
| YEAR | COUNT OF BLANKS |
| 1999 | 7640 |
| 2000 | 7542 |
| 2001 | 6978 |
| 2002 | 7309 |
| 2003 | 7182 |
| 2004 | 7844 |
| 2005 | 7913 |
| 2006 | 8209 |
| 2007 | 8671 |
| 2008 | 7729 |
| 2009 | 7528 |
| 2010 | 7022 |
| 2011 | 6944 |
| 2012 | 6593 |
| 2013 | 5475 |
| 2014 | 5030 |
| 2015 | 4845 |
| 2016 | 1184 |

## MORTGAGE INSURANCE PERCENTAGE (MI%):

|  |  |  |  |
| --- | --- | --- | --- |
| MORTGAGE INSURANCE PERCENTAGE (MI %) | COUNT | YEAR | Percentage |
|  | 9026 | 1999 | 18.052 |
| 0 | 21885 | 1999 | 43.77 |
|  | 44 | 2000 | 0.088 |
| 0 | 32764 | 2000 | 65.528 |
|  | 58 | 2001 | 0.116 |
| 0 | 36990 | 2001 | 73.98 |
|  | 11 | 2002 | 0.022 |
| 0 | 38304 | 2002 | 76.608 |
|  | 10 | 2003 | 0.02 |
| 0 | 40083 | 2003 | 80.166 |
|  | 9 | 2004 | 0.018 |
| 0 | 40437 | 2004 | 80.874 |
|  | 57 | 2005 | 0.114 |
| 0 | 43136 | 2005 | 86.272 |
| 0 | 43086 | 2006 | 86.172 |
| 0 | 39839 | 2007 | 79.678 |
| 0 | 40958 | 2008 | 81.916 |
| 0 | 46460 | 2009 | 92.92 |
| 0 | 46266 | 2010 | 92.532 |
| 0 | 44985 | 2011 | 89.97 |
| 0 | 43711 | 2012 | 87.422 |
| 0 | 40459 | 2013 | 80.918 |
| 0 | 36478 | 2014 | 72.956 |
| 0 | 37309 | 2015 | 74.618 |
| 0 | 9481 | 2016 | 18.962 |

* 1. Zero means No Mortgage insurance
  2. Blanks Means either less than 1% or greater than 55%, so the replacement cannot be generalized in this case. Also, such cases are ~18% in 1999 and ~0.01% in until 2005 and 0 in the later years.
  3. Deriving a new column for mortgage insurance flag is done, where the value is kept No if MI% is zero, otherwise it is made Yes

## NUMBER OF UNITS:

|  |  |  |
| --- | --- | --- |
|  | 1 | 2000 |
|  | 7 | 2004 |

* 1. No missing values for most sample files. Only 1 in the year 2000 and 7 cases in 2004 where number of units is missing
  2. Replaced it with the mode OR Discard the row

## OCCUPANCY STATUS:

* 1. No missing values in the sample files.
  2. Handled the missing value by replacing it by mode or discarding the rows

## ORIGINAL COMBINED LOAN-TO-VALUE(CLTV):

|  |  |  |
| --- | --- | --- |
| ORIGINAL COMBINED LOAN-TO-VALUE (CLTV) | COUNT | Year |
|  | 0 | 1999 |
|  | 3 | 2000 |
|  | 2 | 2001 |
|  | 4 | 2002 |
|  | 3 | 2003 |
|  | 3 | 2004 |
|  | 6 | 2005 |
|  | 1 | 2006 |
|  | 2 | 2007 |
|  | 0 | 2008 |
|  | 0 | 2009 |
|  | 1 | 2010 |
|  | 0 | 2011 |
|  | 2 | 2012 |
|  | 2 | 2013 |
|  | 1 | 2014 |
|  | 2 | 2015 |
|  | 0 | 2016 |

* 1. ~0.01% missing values in the sample files.
  2. If the LTV is less than 80 or greater than 200 or unknown, then this column is unknown. Also if CLTV is less than LTV then, CLTV is set to unknown.
  3. This value is dependent on each individual case, so may not be replaced by mean, median or mode.

## ORIGINAL DEBT-TO-INCOME (DTI) RATIO:

* 1. Ratio greater than 65% are represented as spaces. We replaced it by 70.
  2. Unknowns are represented by null, which we replaced by the median.

## ORIGINAL UPB:

* 1. No missing values in the sample files
  2. If value is missing then discard the rows.

## ORIGINAL LOAN-TO-VALUE:

* 1. Ratios below 6% and greater than 105% are unknown.

|  |  |  |
| --- | --- | --- |
| ORIGINAL LOAN-TO-VALUE (LTV) | COUNT | YEAR |
|  | 0 | 1999 |
|  | 2 | 2000 |
|  | 1 | 2001 |
|  | 1 | 2002 |
|  | 3 | 2003 |
|  | 3 | 2004 |
|  | 6 | 2005 |
|  | 1 | 2006 |
|  | 2 | 2007 |
|  | 0 | 2008 |
|  | 0 | 2009 |
|  | 1 | 2010 |
|  | 0 | 2011 |
|  | 2 | 2012 |
|  | 2 | 2013 |
|  | 1 | 2014 |
|  | 2 | 2015 |
|  | 0 | 2016 |

* 1. Close to zero percent of such occurrence. But, replacing of the values with mean/median cannot be justified as it is specifically said that these values are either less than 6 or greater than 105. So, discarding such rows.

## ORIGINAL INTEREST RATE:

* 1. No missing values
  2. If value is missing then replace by median

## CHANNEL:

* 1. No missing values in sample files
  2. If values are missing then replace by mode

## PREPAYMENT PENALTY MORTGAGE (PPM) FLAG:

|  |  |  |
| --- | --- | --- |
| PREPAYMENT PENALTY MORTGAGE (PPM) FLAG | COUNT | YEAR |
|  | 1247 | 1999 |
|  | 236 | 2000 |
|  | 122 | 2001 |
|  | 171 | 2002 |
|  | 198 | 2003 |
|  | 73 | 2004 |
|  | 49 | 2005 |
|  | 65 | 2006 |
|  | 113 | 2007 |
|  | 1039 | 2008 |
|  | 317 | 2009 |
|  | 336 | 2010 |
|  | 580 | 2011 |
|  | 39 | 2012 |
|  | 4 | 2013 |
|  | 11 | 2014 |
|  | 41 | 2015 |
|  | 7 | 2016 |

* 1. Most number of blanks (unknown) in the year 1999 -> 2.49%, 2008 -> 2.078%

|  |  |
| --- | --- |
| **1999** | **48753** |
| N | 48491 |
| Y | 262 |
| **2000** | **49764** |
| N | 49737 |
| Y | 27 |
| **2001** | **49878** |
| N | 49867 |
| Y | 11 |
| **2002** | **49829** |
| N | 49784 |
| Y | 45 |
| **2003** | **49802** |
| N | 49652 |
| Y | 150 |
| **2004** | **49927** |
| N | 49752 |
| Y | 175 |

* 1. Maximum are “N” throughout the years. 97.5% in 1999, 99.5% in 2000…
  2. We are replacing unknown(blanks) values by mode as it wouldn’t affect the distribution.

1. PRODUCT TYPE:
   1. No missing values found in the observations
   2. If there are any missing values, then it is replaced with “FRM”
2. PROPERTY STATE:
   1. No missing values found in the observations
   2. If there are any missing values, then it is replaced with “Unknown”

## PROPERTY TYPE:

|  |  |  |
| --- | --- | --- |
| PROPERTY TYPE | COUNT | YEAR |
|  | 8 | 2000 |
|  | 11 | 2001 |
|  | 3 | 2002 |
|  | 14 | 2004 |

|  |  |  |
| --- | --- | --- |
| PROPERTY TYPE | COUNT | YEAR |
|  | 8 | 2000 |
| CO | 4090 | 2000 |
| CP | 74 | 2000 |
| LH | 15 | 2000 |
| MH | 244 | 2000 |
| PU | 6531 | 2000 |
| SF | 39038 | 2000 |
|  | 11 | 2001 |
| CO | 3546 | 2001 |
| CP | 45 | 2001 |
| LH | 22 | 2001 |
| MH | 181 | 2001 |
| PU | 5470 | 2001 |
| SF | 40725 | 2001 |
|  | 3 | 2002 |
| CO | 3399 | 2002 |
| CP | 48 | 2002 |
| LH | 12 | 2002 |
| MH | 274 | 2002 |
| PU | 5053 | 2002 |
| SF | 41211 | 2002 |
|  | 14 | 2004 |
| CO | 3616 | 2004 |
| CP | 210 | 2004 |
| LH | 35 | 2004 |
| MH | 529 | 2004 |
| PU | 6829 | 2004 |
| SF | 38767 | 2004 |

* 1. No missing values for most of the years.
  2. Very few missing values observed for years 2000, 2001, 2002 and 2004.
  3. Replaced the missing values with the mode (“SF” as observed) because most number of records are categorized as Single Family Home (77% to 82%)

## POSTAL CODE:

|  |  |  |
| --- | --- | --- |
| POSTAL CODE | COUNT | YEAR |
|  | 1 | 1999 |
|  | 72 | 2000 |
|  | 1 | 2001 |
|  | 1 | 2002 |
|  | 0 | 2003 |
|  | 0 | 2004 |
|  | 1 | 2005 |
|  | 0 | 2006 |
|  | 0 | 2007 |
|  | 0 | 2008 |
|  | 0 | 2009 |
|  | 0 | 2010 |
|  | 0 | 2011 |
|  | 0 | 2012 |
|  | 0 | 2013 |
|  | 0 | 2014 |
|  | 0 | 2015 |
|  | 0 | 2016 |

* 1. 72 of 50000 unknowns in 2000, 1 row each in 1999, 2001,2002 and 2005 of unknowns
  2. Replaced the blanks with 99999 as unknown value
  3. Future Scope: Get a complete dictionary of Metropolitan Statistical Area or Metropolitan Division codes and map the MSA or MD for the row to the dictionary to find the missing postal code

## LOAN SEQUQENCE NUMBER:

* 1. Unique Identifier Column.
  2. No missing values. If the value is missing for a row, then replace by random Loan sequence number the complete row or generating a unique identifier UUID
  3. Derived two new columns for origination year and origination quarter

## LOAN PURPOSE:

* 1. No missing values in the sample files.
  2. If the values are missing then, loan purpose is unknown. Assuming that the percentage of such occurrence in the yearly data would be close (if not equal to) 0%, and it wouldn’t affect the distribution of the data, we replaced it by the mode of the column

## ORIGINAL LOAN TERM:

* 1. No missing values observed.

## NUMBER OF BORROWERS:

|  |  |  |
| --- | --- | --- |
| NUMBER OF BORROWERS | COUNT | YEAR |
|  | 30 | 1999 |
|  | 20 | 2000 |
|  | 11 | 2001 |
|  | 9 | 2002 |
|  | 7 | 2003 |
|  | 14 | 2004 |
|  | 17 | 2005 |
|  | 17 | 2006 |
|  | 23 | 2007 |
|  | 19 | 2008 |
|  | 6 | 2009 |
|  | 0 | 2010 |
|  | 0 | 2011 |
|  | 0 | 2012 |
|  | 0 | 2013 |
|  | 0 | 2014 |

* 1. 0% to 0.6% Missing values found.
  2. Replacing missing values with the mode.

## SELLER NAME:

* 1. No missing values found in the sample files.
  2. Replacing missing values by “Unknown”

## SERVICES NAME:

* 1. No missing values found in the sample files.
  2. Replacing missing values by “Unknown”

## SUPER CONFORMING FLAG:

|  |  |  |
| --- | --- | --- |
| SUPER CONFORMING FLAG | COUNT | YEAR |
| Y | 80 | 2008 |
| Y | 1236 | 2009 |
| Y | 1364 | 2010 |
| Y | 1967 | 2011 |
| Y | 2189 | 2012 |
| Y | 1718 | 2013 |
| Y | 1995 | 2014 |
| Y | 2223 | 2015 |
| Y | 492 | 2016 |

1. Per the data dictionary, all the missing values are Not super conforming, so replaced the missing values by “N”

# PERFORMANCE FILE

1. LOAN SEQUENCE NUMBER:
   1. Derived two new columns for origination year and origination quarter
2. MONTHLY REPORTING PERIOD:
   1. Derived two new columns for monthly reporting period year and month
3. CURRENT ACTUAL UPB:
4. CURRENT LOAN DELINQUENCY STATUS:
   1. No Missing values observed in the sample files.
   2. Replacing missing values with “XX” which is also used for unknown.
5. LOAN AGE:
   1. No missing values observed.
6. REMAINING MONTHS TO LEGAL MATURITY:
   1. No missing values found.
7. REPURCHASE FLAG:
   1. This field is only populated at loan termination. For all others the value is not applicable.
   2. Replacing nulls with NA.
8. MODIFICATION FLAG:
   1. Replacing nulls with “NO” (Not modified)
9. ZERO BALANCE CODE:
   1. Replacing nulls and spaces with “NA” as it is not applicable if the balance is not reduced to zero.
10. ZERO BALANCE EFFECTIVE DATE:
    1. Replacing missing values with 999999, which will denote not applicable.
    2. Deriving 2 new columns for zero balance effective year and month.
11. CURRENT INTEREST RATE:
    1. Replacing empty values with 0.
12. DUE DATE OF LAST PAID INSTALLMENT:
    1. Replacing missing values with 999999.
    2. Deriving 2 new columns for due year and month of last paid installment.
13. Replacing missing values with 0 for the following columns
    1. MI RECOVERIES
    2. NET SALES PROCEEDS
    3. NON MI RECOVERIES
    4. EXPENSES
    5. LEGAL COSTS
    6. MAINTENANCE AND PRESERVATION COSTS:
    7. TAXES AND INSURANCE:
    8. MISCELLENEOUS EXPENSES:
    9. ACTUAL LOSS CALCULATION:­
    10. MODIFICATION COST
    11. CURRENT DEFERRED UPB

# CLASSIFICATION (LOGISTIC REGRESSION)

## summary of the logistic regression

> summary(modelLogit)

Call:

glm(formula = DELINQUENT ~ ., family = binomial(link = "logit"),

data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.4397 -0.3161 -0.2212 -0.1563 3.5311

Coefficients: (6 not defined because of singularities)

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

(Intercept) -1.289e+00 5.889e-02 -21.882 < 2e-16 \*\*\*

CURRENTACTUALUPB 3.625e-06 3.433e-08 105.574 < 2e-16 \*\*\*

LOANAGE 2.221e-02 8.232e-05 269.852 < 2e-16 \*\*\*

REMAININGMONTHSTOLEGALMATURITY 2.346e-03 4.261e-05 55.061 < 2e-16 \*\*\*

ZEROBALANCECODE -4.294e-03 2.531e-04 -16.964 < 2e-16 \*\*\*

CURRENTINTERESTRATE 2.093e-02 4.267e-03 4.905 9.36e-07 \*\*\*

MONTHLYREPORTINGYEAR NA NA NA NA

MONTHLYREPORTINGMONTH NA NA NA NA

REPURCHASEFLAGYES 2.890e+00 1.895e-01 15.255 < 2e-16 \*\*\*

MODIFICATIONFLAGYES 5.272e+00 1.060e-01 49.717 < 2e-16 \*\*\*

ZEROBALANCEEFFECTIVEYEAR NA NA NA NA

ZEROBALANCEEFFECTIVEMONTH NA NA NA NA

DUEDATEOFLASTPAIDINSTALLMENTYEAR NA NA NA NA

DUEDATEOFLASTPAIDINSTALLMENTMONTH NA NA NA NA

CREDIT\_SCORE -1.025e-02 3.998e-05 -256.290 < 2e-16 \*\*\*

NUMBER\_OF\_UNITS -1.825e-01 1.235e-02 -14.777 < 2e-16 \*\*\*

ORIGINAL\_INTEREST\_RATE 5.930e-01 8.312e-03 71.348 < 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: 1354776 on 3551480 degrees of freedom

Residual deviance: 1155220 on 3551470 degrees of freedom

AIC: 1155242

Number of Fisher Scoring iterations: 6

> pR2(modelLogit)

llh llhNull G2 McFadden r2ML r2CU

-5.776098e+05 -6.773880e+05 1.995563e+05 1.472984e-01 5.464012e-02 1.722893e-01

## CONFUSION MATRIX for LOGISTIC REGRESSION

> confusionMatrix(data=factor(pred.resp.level),reference=factor(test$DELINQUENT),positive='1')

Confusion Matrix and Statistics

## prediction

|  |
| --- |
| Reference |
| Prediction 0 1 |
| 0 2665523 209453 |
| 1 6246 5076 |

Accuracy : 0.9253

95% CI : (0.925, 0.9256)

No Information Rate : 0.9257

P-Value [Acc > NIR] : 0.9957

Kappa : 0.0378

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.023661

Specificity : 0.997662

Pos Pred Value : 0.448331

Neg Pred Value : 0.927146

Prevalence : 0.074327

Detection Rate : 0.001759

Detection Prevalence : 0.003923

Balanced Accuracy : 0.510662

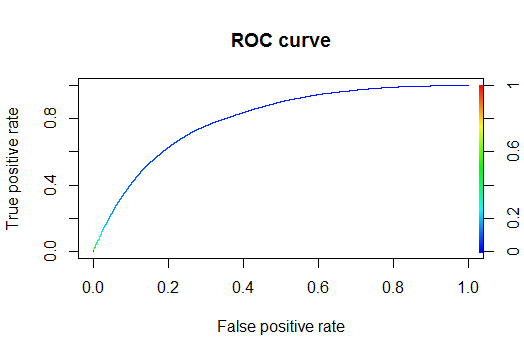
'Positive' Class : 1

> delinquent.logistic.error

[1] 0.07473206

>

## ROC CURVE



## NEURAL NETWORK

fitnn <- nnet(DELINQUENT ~ ., traindatannet, size=20,

+ maxit = 90, entropy = TRUE, softmax = FALSE,censored = FALSE, skip = FALSE,

+ rang = 0.7, Hess = FALSE, trace = TRUE, MaxNWts = 1000, abstol = 1.0e-4,

+ decay = 15e-4, reltol = 1.0e-8, hidden = 2,threshold = 0.01,act.fct="tanh")

# weights: 361

initial value 30046.231065

iter 10 value 10542.487357

iter 20 value 10494.425353

iter 30 value 10461.239244

iter 40 value 10452.427974

iter 50 value 10450.919183

iter 60 value 10450.736090

iter 70 value 10450.191369

iter 80 value 10449.676547

iter 90 value 10448.726339

…

…

..

final value 10448.726339

> fitnn

a 16-20-1 network with 361 weights

inputs: CURRENTACTUALUPB LOANAGE REMAININGMONTHSTOLEGALMATURITY ZEROBALANCECODE CURRENTINTERESTRATE MONTHLYREPORTINGYEAR MONTHLYREPORTINGMONTH REPURCHASEFLAGYES MODIFICATIONFLAGYES ZEROBALANCEEFFECTIVEYEAR ZEROBALANCEEFFECTIVEMONTH DUEDATEOFLASTPAIDINSTALLMENTYEAR DUEDATEOFLASTPAIDINSTALLMENTMONTH CREDIT\_SCORE NUMBER\_OF\_UNITS ORIGINAL\_INTEREST\_RATE

output(s): DELINQUENT

options were - entropy fitting decay=0.0015

## CONFUSION MATRIX for NEURAL NETWORK

> confusionMatrix(data=pred.resp.nnet.factor,reference=factor(testdatannet$DELINQUENT), positive='1')

Confusion Matrix and Statistics

## PREDICTION

|  |
| --- |
| Reference |
| Prediction 0 1 |
| 0 26 140 |
| 1 46450 2764 |

Accuracy : 0.0682

95% CI : (0.066, 0.0704)

No Information Rate : 0.9419

P-Value [Acc > NIR] : 1

Kappa : -0.0041

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.95179

Specificity : 0.01372

Pos Pred Value : 0.05616

Neg Pred Value : 0.82188

Prevalence : 0.05808

Detection Rate : 0.05528

Detection Prevalence : 0.98428

Balanced Accuracy : 0.48275

'Positive' Class : 1

> #computing the overall error - 0.32

> delinquent.neural.error <- 1- sum(pred.resp.nnet.factor==testdatannet$DELINQUENT)/length(testdatannet$DELINQUENT)

> delinquent.neural.error

[1] 0.9318

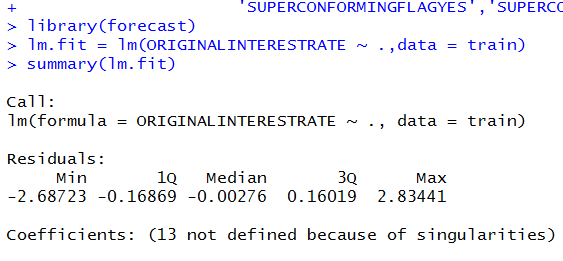
# PREDICTION

Regression model for the interest rate.

### Linear regression

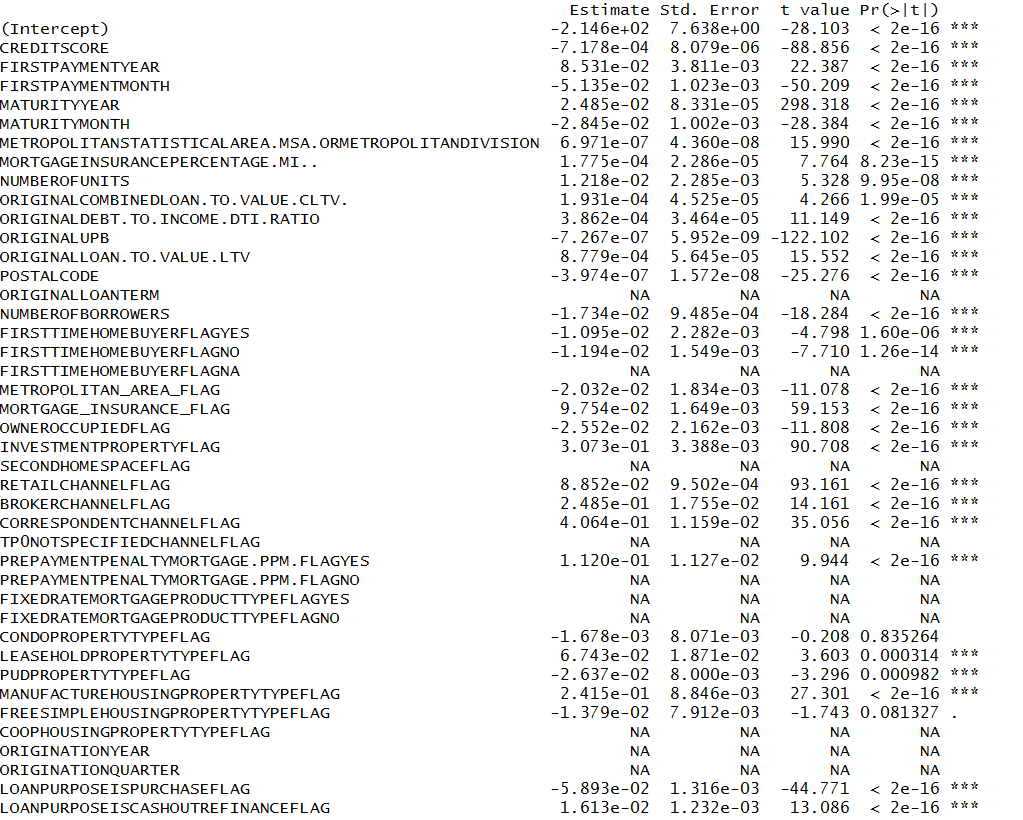
#### All variables CONSIDERED (45)

The dataset ‘train’ consists of all the columns from the cleaned sample origination files.

OUTPUT:

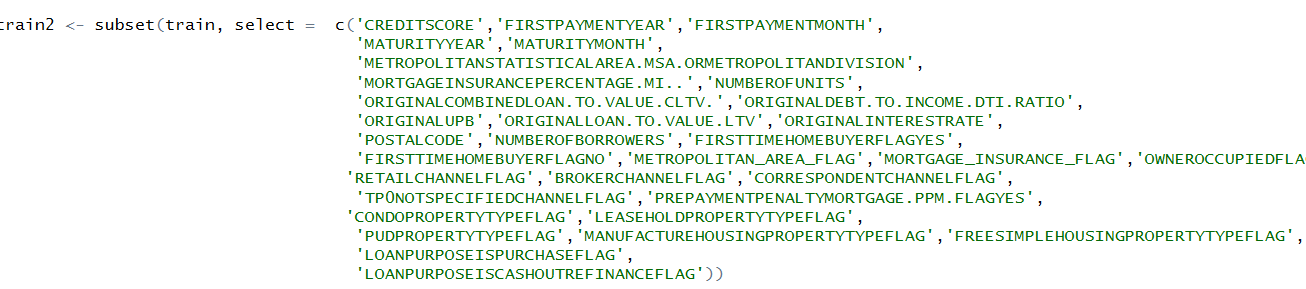
The coefficients are given for the columns:

13 columns were removed as they were singular i.e, removing these columns will not affect the coefficients of the model

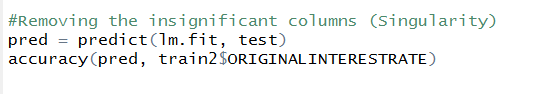


## PREDICTION ON TEST DATA

#### MODEL two – SELCTing statistically significant columns (p <0.05) –

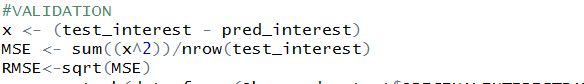
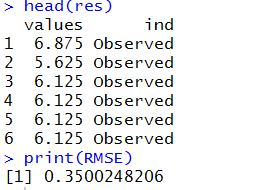
Select the columns that were considered significant by the previous model.

## PREDICTION ON TEST DATA

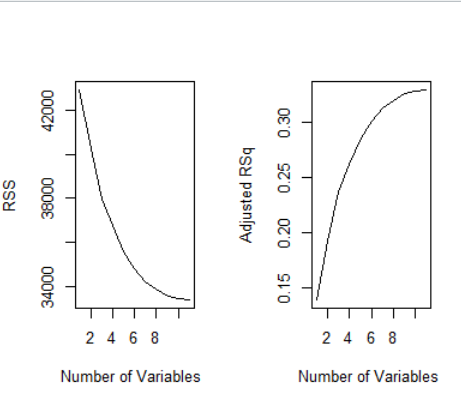


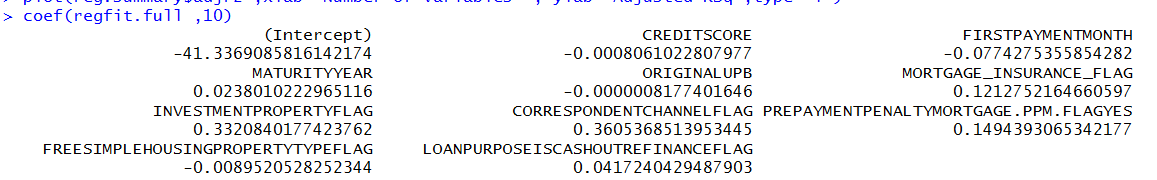
The accuracy measure do not change as expected

## validation with observed data

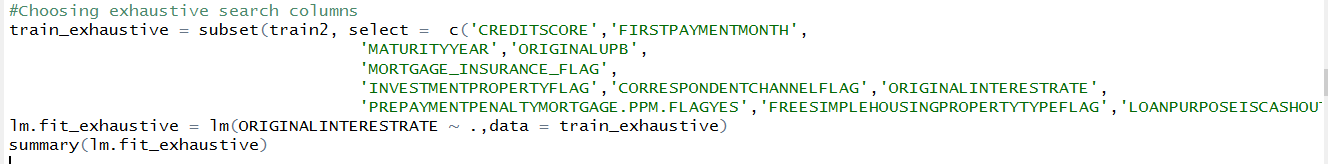


## eXHAUSTIVE SEARCH



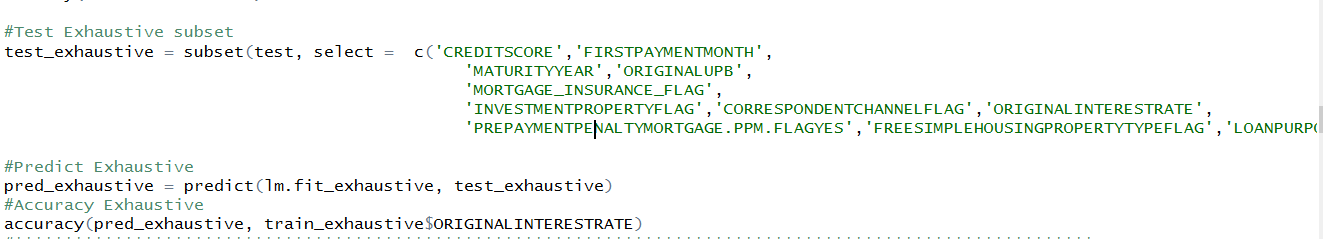
Since RSS seems to flatten at about 10 variables, we can conclude that these ten variables will influence my model the most. We obtain the 10 columns and their corresponding co-efficients.

## training on the columns selected using exhaustive search

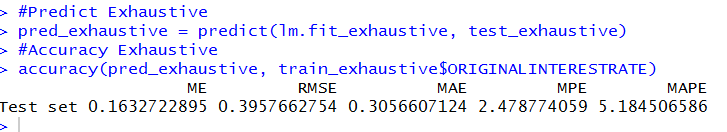


## Summary of the model

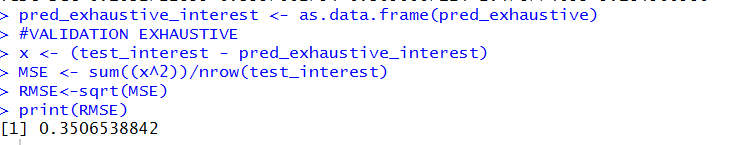
## predicting



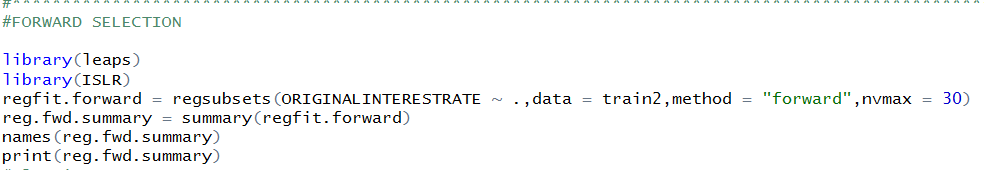
## results



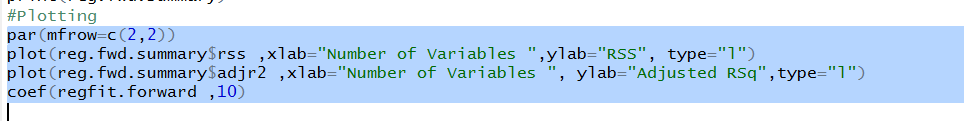
## validation

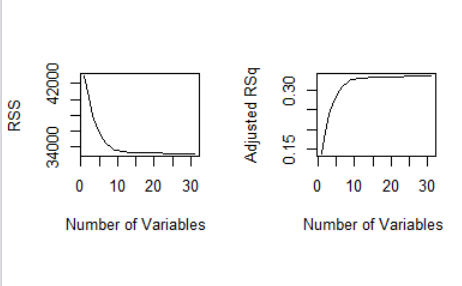


## forward selection



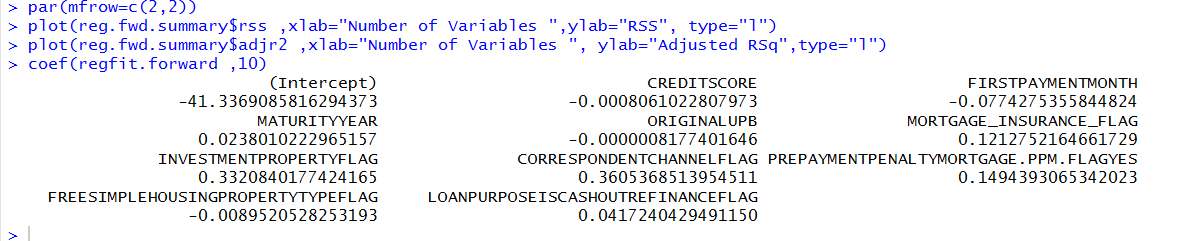
## Checking curves for column selection



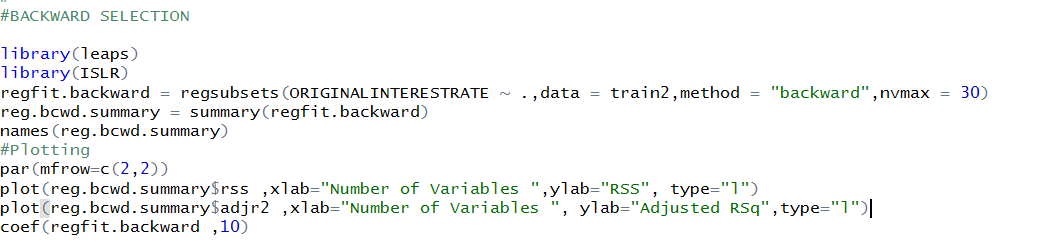


RSS flattens out at 10 variables. This means that we don’t need more than ten significance columns to explain the model

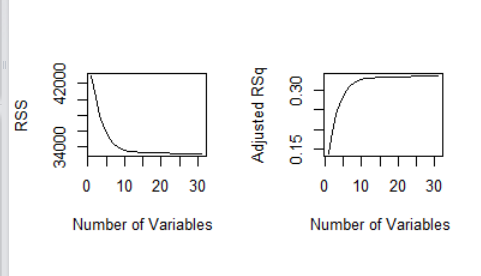
We move ahead to pick these ten columns and their co-efficients

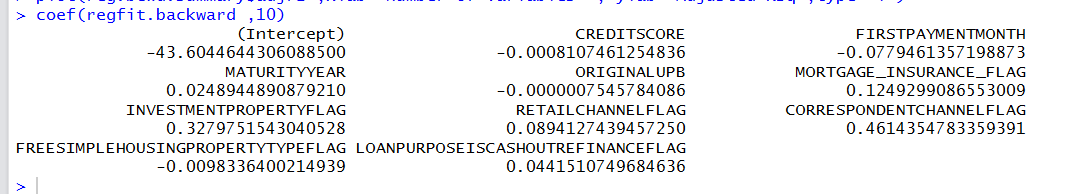
These columns are exactly the same as we got for exhaustive search. Hence, we expect the same accuracy

## BACKWARD selection

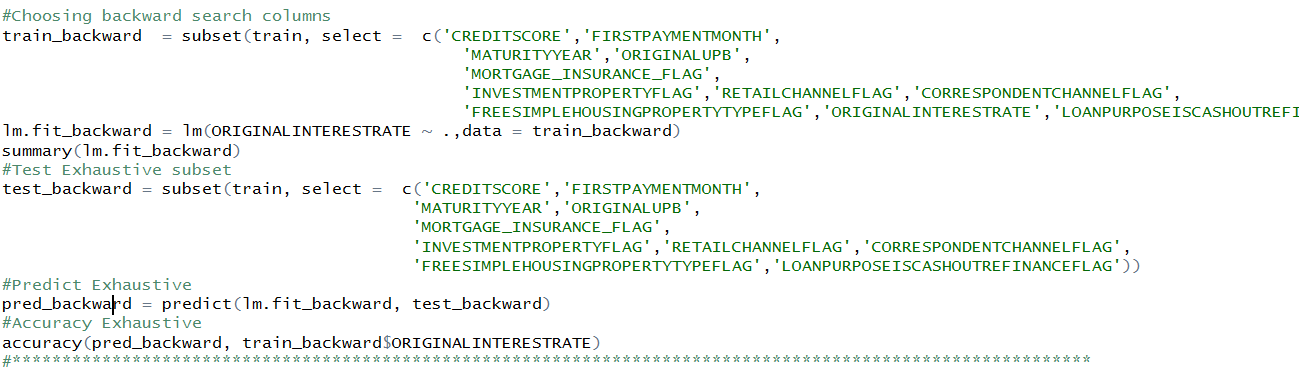


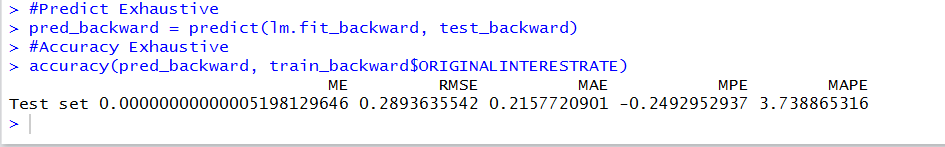
A similar curve, flattening at 10 variables is observed.

And the following ten varaibles are chosen

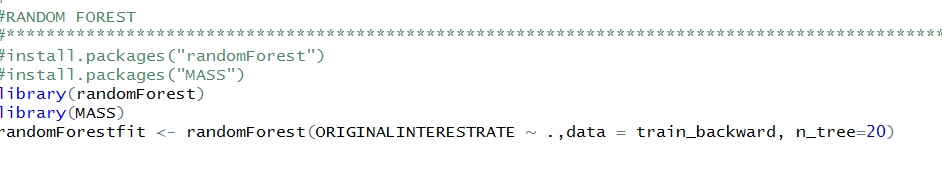


## Subset columns as per backward selection and prediction





## Algorithm 2 – random forest

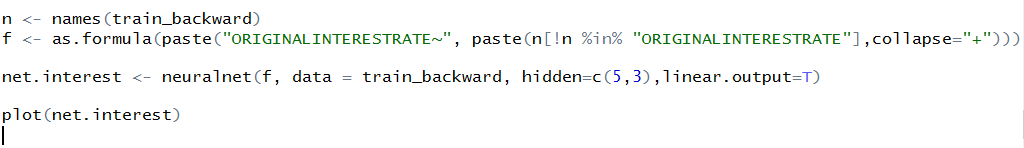
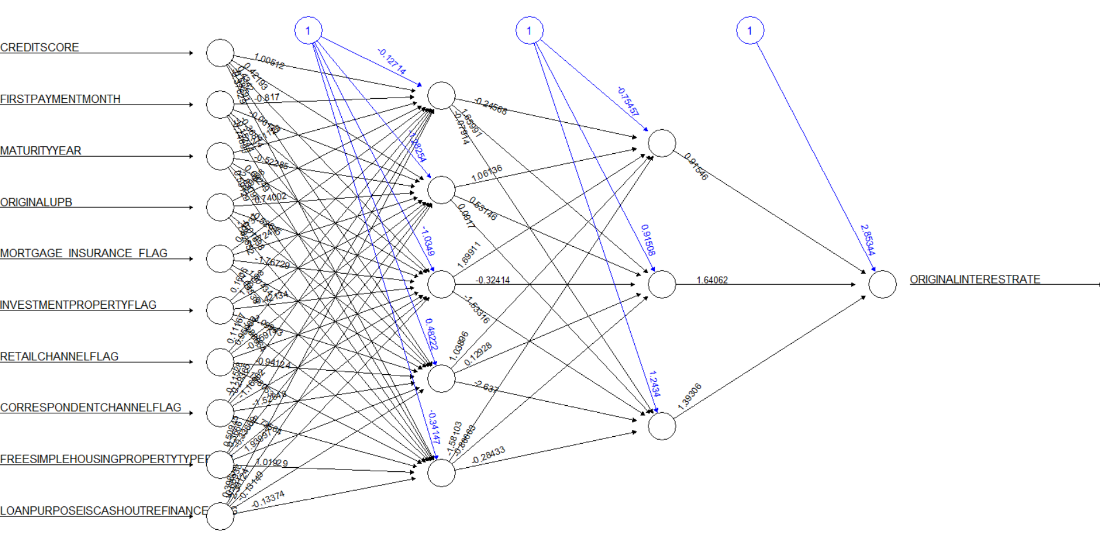


## prediction

### C:\Users\Sneha Ravi\AppData\Local\Microsoft\Windows\INetCacheContent.Word\image.png

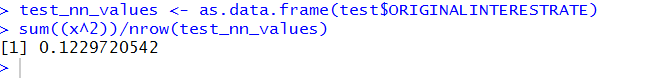
## algorithm 3- neural network

## creating the network



## prediction

## check mean square error

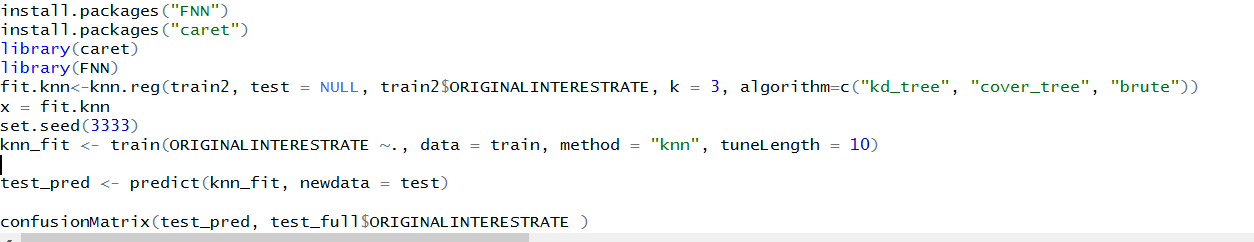


> RMSE = sqrt(0.1220720542)

> print(RMSE)

[1] 0.3506

## aLGORITHM 4 -knn



## comparison of models

The RMSE values across different models , it was evident that the backward selection method gave the best set of variables for a highly accurate model with RMSE of 0.2893 compared to the other models:

|  |
| --- |
| RMSE |
| Linear Regression with all variables 0.3996 |
| Exhaustive and Forward Selection with Linear Regression: 0.3957 |
| Backward Selection 0.2893 |
| Random Forest 0.44 |
| Neural Network 0.35 |