Housing Loan Prediction

Presented by Group 12 :

Keyuri Bhuwariya

Shweta Patel

Utsav Fadia

TejasKumar Patel

Tej Kurani

Project Overview

People approach banks to take loans to fulfil their needs accordingly. This practice has been increasing day by day across all sectors of the society such as education purposes, business purposes and predominantly for agriculture. But some people take this advantage of taking the loan and misuse that money for some other purposes and do not re-pay the amount back to the loan. There are high chances that the bank might lose its money in case if loan sanctioned to such customers. With technology developing very rapidly these days, data mining plays a key role to solve such issues. This Projects targets to predict and classify the customers who will be able to repay the money back to the bank. We use Classification task for the machine learning algorithms for predictive modelling in such scenarios.

Business Understanding

Bargaining Power of Suppliers:

- Capital is the primary resource on any bank and there are four major suppliers :

1. Customer deposits.
2. Mortgages and loans.
3. Mortgage-backed securities.
4. Loans from other financial institutions.

- By utilizing these four major suppliers, the bank can be sure that they have the necessary resources required to service their customers' borrowing needs while maintaining enough capital to meet withdrawal expectations.
- The power of the suppliers is largely based on the market, their power is often considered to fluctuate between medium to high.

Bargaining Power of Customers:

- The individual doesn't pose much of a threat to the banking industry, but one major factor affecting the power of buyers is relatively high switching costs.
- To try and convince customers to switch to their bank they will often lower the price of switching, though most people still prefer to stick with their current bank.
- The internet has greatly increased the power of the consumer in the banking industry and reduced the cost for consumers to compare the prices of opening accounts as well as the rates offered at various banks.
- ING Direct introduced high yield savings accounts to catch the buyers' attention, they went a step further and made it very easy for customers to transfer their money from their current bank to ING.

Threat of New Entrants:

- With so many new banks entering the market each year the threat of new entrants should be extremely high. However, due to mergers and bank failures the average number of total banks decreases by roughly 253 a year.
- Because the industry deals with other people's money and financial information new banks find it difficult to start up. Due to the nature of the industry people are more willing to place their trust in big name, well known, major banks who they consider to be trustworthy.
- The banking industry has undergone a consolidation in which major banks seek to serve a customer’s financial needs. This consolidation furthers the role of trust as a barrier to entry for new banks looking to compete with major banks, as consumer are more likely to allow one bank to hold all their accounts and service their financial needs.

Availability of Substitute Products:

- Some of the banking industry's largest threats of substitution are not from rival banks but from non-financial competitors.
- The industry does not suffer any real threat of substitutes as far as deposits or withdrawals; however, insurances, mutual funds, and fixed income securities are some of the many banking services that are also offered by non-banking companies.
- There are also the threat of payment method substitutes and loans are relatively high for the industry. For example, big name electronics, jewellers, car dealers, and more tend to offer preferred financing on "big ticket" items. The non-banking companies offer a lower interest rate on payments then the consumer would otherwise get from a traditional bank loan.

Intensity of Competitive Rivalry in the Industry:

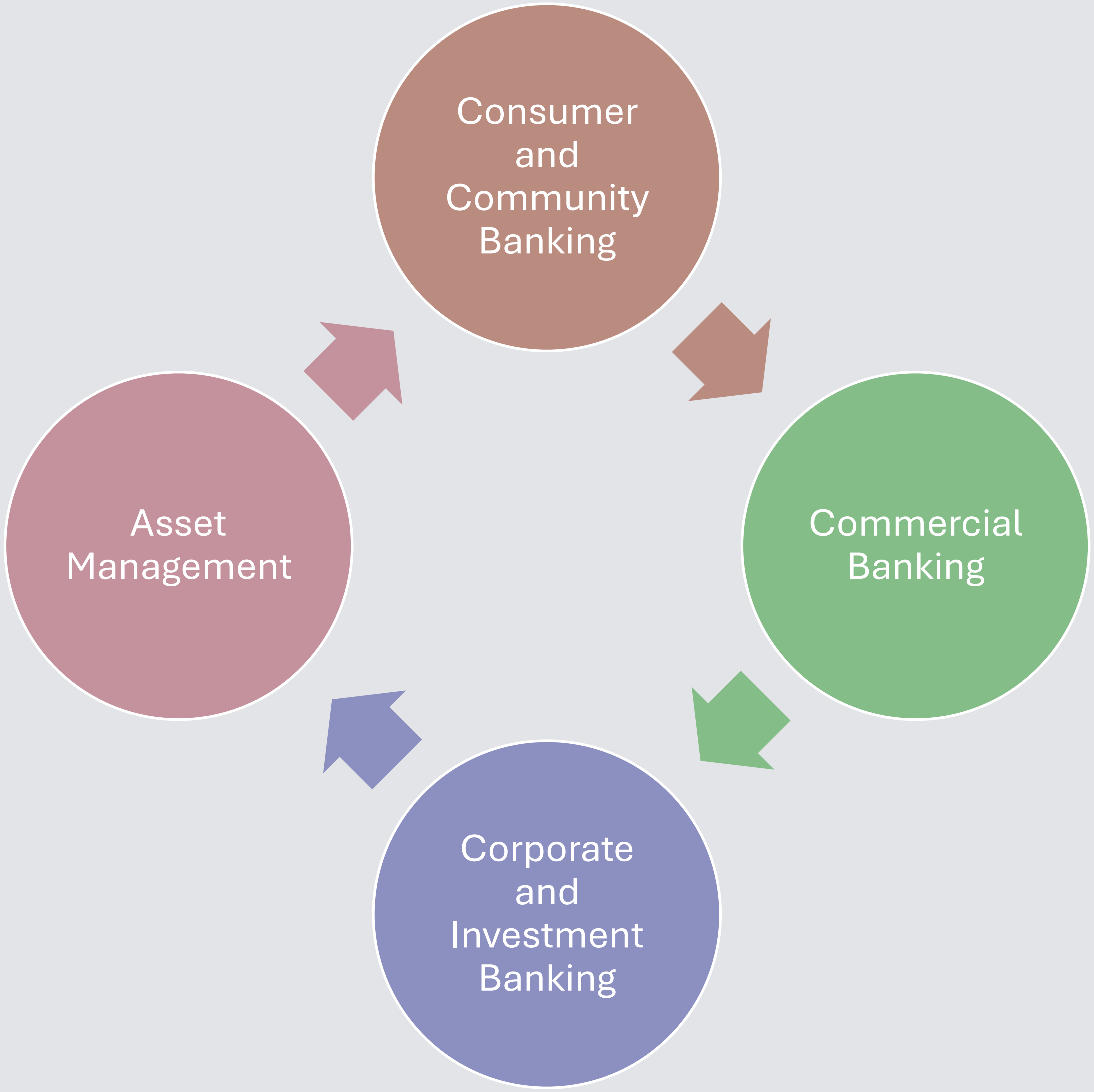
- The banking industry is considered highly competitive. The financial services industry has been around for hundreds of years, and just about everyone who needs banking services already has them. Because of this, banks must attempt to lure clients away from competitor banks. They do this by offering lower financing, higher rates, investment services, and greater conveniences than their rivals.
- The banking competition is often a race to determine which bank can offer both the best and fastest services, but has caused banks to experience a lower ROA (Return on Assets). Given the nature of the industry it is more likely to see further consolidation in the banking industry. Major banks tend to prefer to acquire or merge with other banks than to spend money marketing and advertising.

Firm Description:

JPMorgan Chase (JPM) is a major global bank holding and financial services company. It is a universal banking company that provides commercial, retail, and investment banking services. It is one of the four principal money centre banks in the United States, along with Wells Fargo, Bank of America, and Citigroup. With more than $2.3 trillion in assets, JPMorgan is one of the 10 largest banks worldwide.

The company, as we know it today, is the result of a series of mergers of a group of major U.S. banks. It is one of the four major banks in the United States, along with Citibank, Bank of America, and Wells Fargo. JPMorgan operates as a bank holding company with several subsidiaries engaged in the company's four main areas of financial enterprise:

Four main areas of Financial enterprise

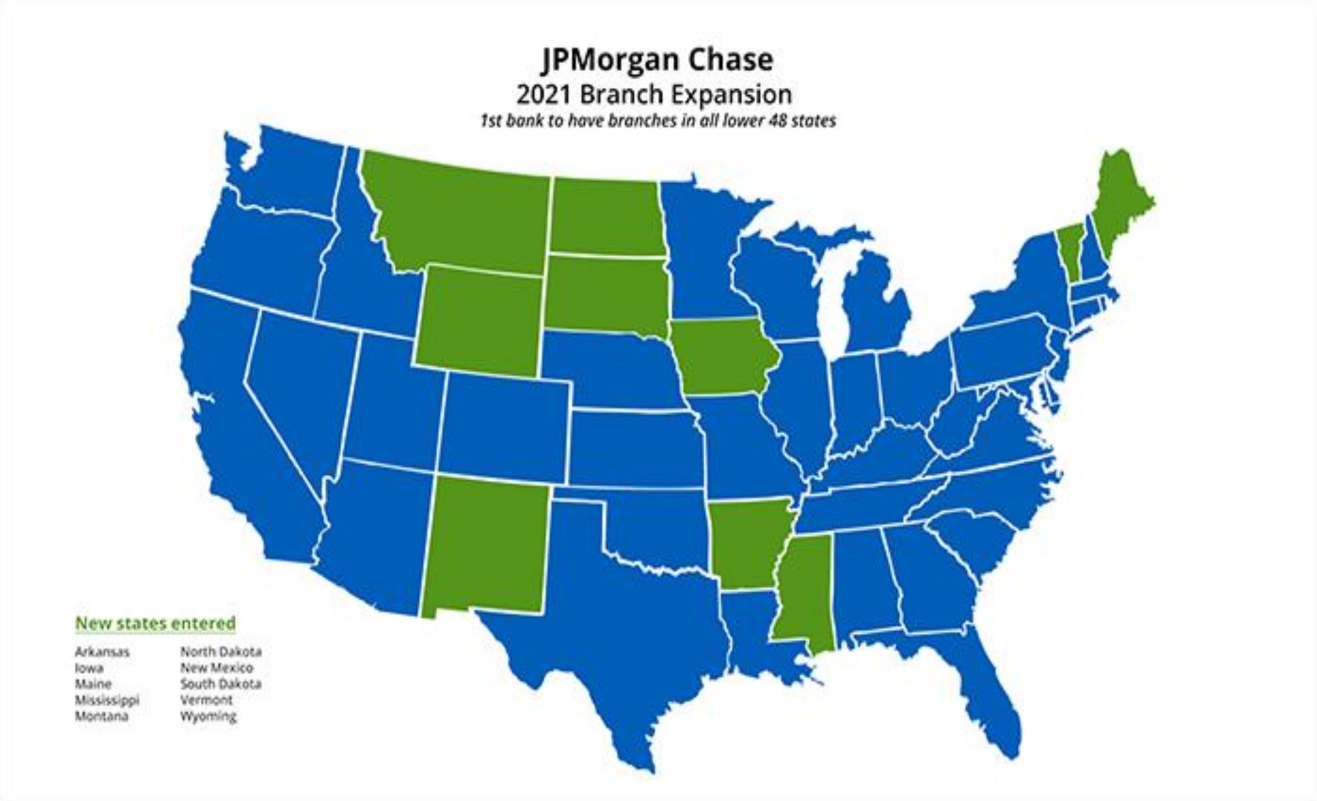


Revenues Tied to each Service

- **Consumer Banking:** Chase's Consumer & Community Banking revenue for 2020 was $51.3 billion. Chase holds numerous #1 rankings within the US consumer banking market.
- **Commercial Banking:** In 2020 Chase's Commercial Banking business delivered net income of $2.6 billion on $9.3 billion in revenue.
- **Corporate and Investment Banking:** In 2020, CIB generated earnings of $17.1 billion on revenue of $49.3 billion, achieving a 20% return on equity.
- **Asset Management:** Net revenue for this segment was $14.2 billion in 2020, an increase of 5%. Net interest income rose 2% to $3.4 billion, driven by higher deposit and loan balances

Geographical Footprints

The number of employees of JPMorgan Chase worldwide increased overall between 2008 and 2021, despite some fluctuations. The number of JPMorgan Chase employees amounted to 271,025 in 2021 and has the geographic footprint over 48 states in USA.



SWOT Analysis:

Strengths

1.Strong brand name and good financial position

2.Global presence and employs over 250,000 around the world

3.Excellent services for customers through extensive retail network

4.Good brand visibility in the B2B segment

5.Largest bank in US in terms of sales, market value, assets and profits

Weaknesses

1.Overdependence on USA

2.Stiff competition from other financial service providers

3.Fluctuating markets result in instability

Opportunities

1.Expansion in other countries

2.Diversifying portfolios for customers

3.Investments across the world

4. Commercial banking, and JVs

Threats

1.Changing govt regulations and financial crisis like recessions

2.Unstable mortgage market

Data Description

This dataset has 29 columns with 1,776 rows, with 28 independent and 1 dependent variable. The dependent variable is **approve (1 = “yes” and 0 = “No”).** The period of time for this dataset is 2 years.

Data Source

Referred by Undergraduate Professor

URL: http://fmwww.bc.edu/ec-p/data/wooldridge/loanapp.des

Dataset

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

loan.r* | loan_approve_project | loan.df

Filter

	occ	loanamt	appinc	unit	married	dep	emp	self	hexp	price	other	liq	gdlin	mortg	pubrec	hratAmt	obratAmt	fixadj	term	cosign	netw	sch	hispan	male	approve	mortno	mortperf	chist	multi
1	1	128	73092	1	1	1	0	0	1391.0	160.0	0.000	5.200e+01	1	2	0	16474.937	24924.37	1	360	0	847.000	1	0	1	0	0	1	1	0
2	1	128	84984	1	0	0	1	0	1371.0	143.0	0.000	3.700e+01	1	2	0	16146.960	22095.84	0	180	0	40.000	1	0	1	1	0	1	1	0
3	1	66	36000	1	1	0	0	1	839.0	110.0	0.000	1.900e+01	1	2	1	8640.000	13320.00	0	360	0	158.000	1	0	1	1	0	1	0	0
4	1	120	59292	1	1	0	0	0	1341.0	134.0	0.000	3.100e+01	1	2	0	14882.292	19032.73	0	360	0	69.000	0	0	1	1	0	1	1	0
5	1	111	62004	1	0	0	0	0	1122.0	138.0	0.000	1.690e+02	1	2	0	13020.840	20461.32	0	360	0	262.000	0	0	1	1	0	1	0	0
6	1	141	72000	1	0	0	0	0	1870.0	157.0	0.000	3.800e+01	1	2	0	23040.000	25920.00	0	360	0	46.000	1	0	1	1	0	1	1	0
7	1	276	90000	1	0	0	0	0	2706.0	346.0	0.000	1.420e+02	1	2	0	32400.000	33300.00	1	360	0	214.000	1	0	1	1	0	1	1	0
8	1	100	69096	1	1	1	1	0	1112.0	99.7	0.000	2.450e+01	1	2	0	13342.438	21212.47	0	360	0	38.800	1	0	1	1	0	1	1	0
9	1	267	115356	1	1	2	0	1	2959.0	422.0	0.000	1.110e+02	1	1	0	30580.876	56524.44	0	360	0	374.000	1	0	1	1	1	0	1	0
10	1	175	96600	1	1	0	0	0	2081.0	346.7	0.000	1.866e+02	1	1	0	25116.000	35742.00	0	180	0	348.400	1	0	1	1	1	0	1	0
11	1	95	46188	1	1	0	0	0	1005.0	135.0	0.000	5.233e+01	1	2	0	12055.068	17135.75	0	360	0	66.032	0	0	1	1	0	1	0	0
12	1	158	66480	1	1	4	0	0	1442.0	210.0	39.000	1.880e+02	0	2	1	19279.200	27921.60	1	360	0	239.000	1	0	1	0	0	1	0	0
13	1	110	68640	1	1	0	0	0	1373.0	143.0	0.000	1.530e+02	1	1	0	16473.600	21964.80	0	360	0	345.000	1	0	1	1	1	0	1	0
14	1	40	40920	1	1	0	0	0	538.0	197.0	0.000	3.400e+01	1	1	0	6547.200	6547.20	1	360	0	155.000	0	0	1	1	1	0	1	0
15	1	70	33600	1	1	1	1	1	688.0	78.0	0.000	1.000e+01	1	2	0	8232.000	12465.60	1	360	0	89.000	1	0	1	1	0	1	1	0
16	1	168	66384	1	1	2	0	0	1733.0	210.0	0.000	6.600e+01	1	2	0	20778.192	21375.65	0	360	0	83.100	1	0	1	1	0	1	1	0
17	2	114	97128	1	1	1	0	0	1378.0	170.0	0.000	3.600e+01	1	1	0	19425.600	25253.28	0	360	0	514.000	1	0	1	1	1	0	1	0
18	1	349	285708	1	1	1	0	0	4575.0	387.0	0.000	5.700e+01	1	1	0	59998.680	77141.16	0	360	0	598.000	1	0	1	1	1	0	1	0

Showing 1 to 18 of 1,776 entries, 29 total columns

Columns Description

Independent Variables	Description
Occ	Customer's Occupancy
Loanamt	Loan amount in thousands
Appinc	Applicants' income in thousands
Unit	No. of units in the property
Married	=1 (Married), =0 (Unmarried)
Dep	No. of Dependents
Emp	No. of years of employment
Self	=1 (Self Employed)
Hexp	Proposing housing expense
Price	Purchase price of the house
Other	Other financing (in thousands)
Liq	Liquid Assets worth

Column Description

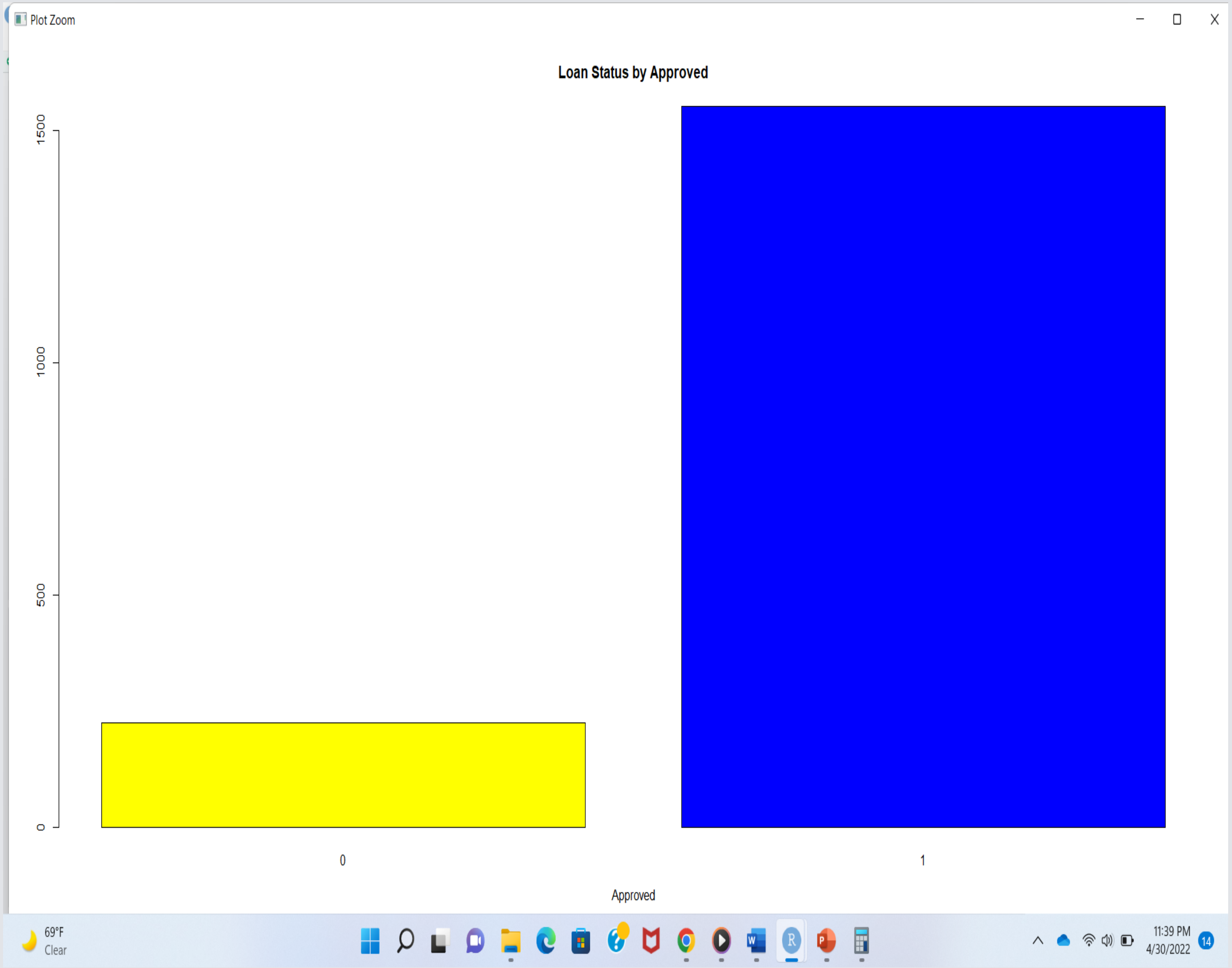
Gdlin	Customer’s credit history meets requirement (=1)
Mortg	Credit History on mortgage payment
Pubrec	=1 (Filed Bankruptcy)
HratAmt	Housing expense w.r.t Total Income in thousands
ObratAmt	Other obligations w.r.t of Total Income in thousands
FixAdj	Fixed or Adjustable rate (=1 Fixed)
Term	Term of Loan in months
Cosign	Is there a co-signer (=1 Yes)
Netw	Net Worth (Liquid Assets – Liabilities)
Sch	=1 (If >12 years of schooling)
Hispan	Is Hispanic (=1)
Male	Is Male (=1)
Mortno	No Mortgage history
Mortperf	No late mortgage payments
Chist	=0, if accounts deliq. >= 60 days
Multi	=1, if two or more units

Data Cleaning

- We have aggregated the sum value of columns Applicant Income and Co App Income into AppIncome.
- Converting hrat and obrat from % to actual amount in thousands (i.e., normalizing)
- Eliminated column loanprc (i.e., amt/price), and considered columns amt and price individually.
- Removing column rep (No. of Credit reports) and cons (Credit history on consumer stuff), as we have column gdlin.

Data Analysis

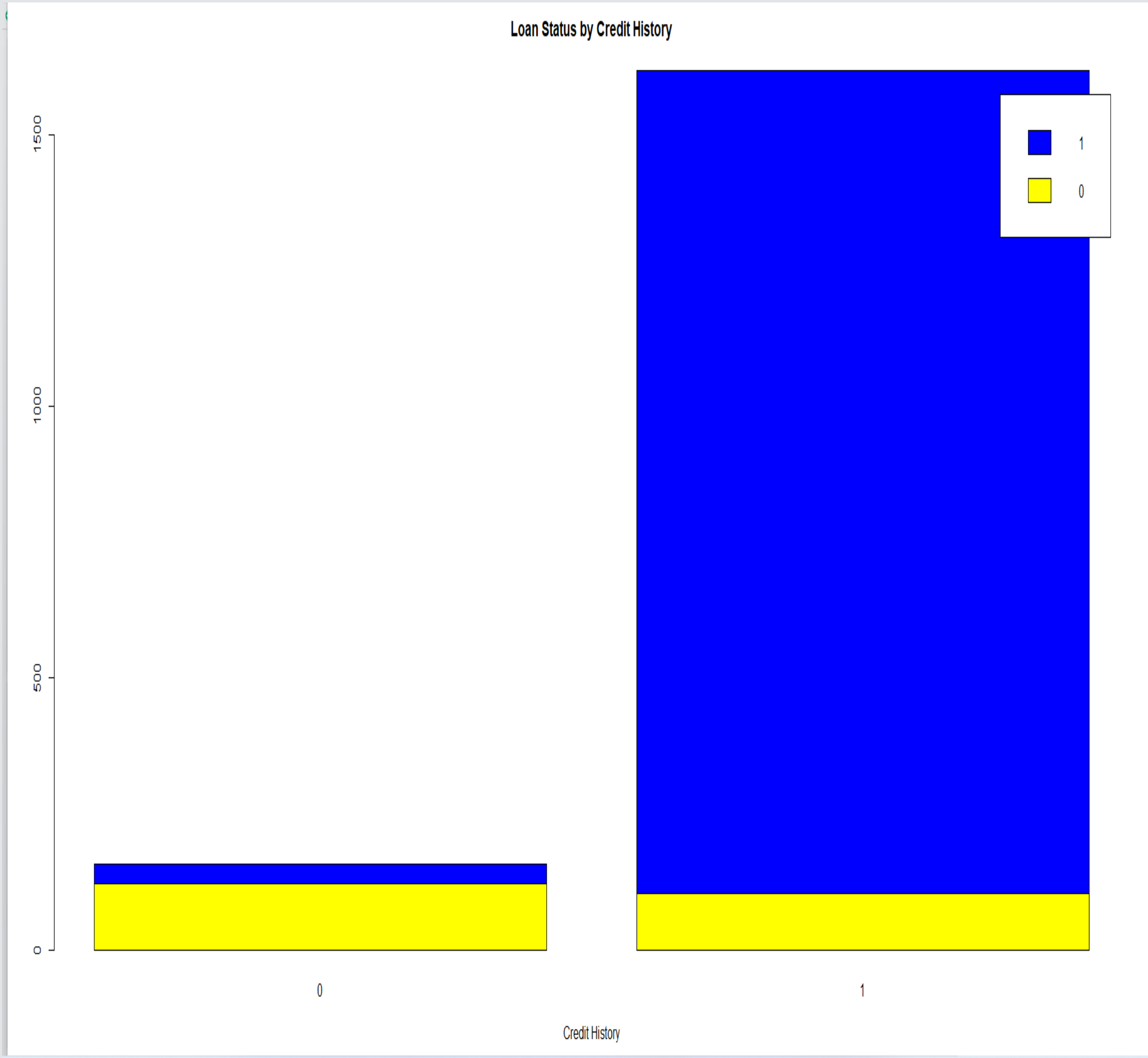
Loan Status – 0 indicates rejected, 1 indicates approved on x axis. We can see that approximately 200 applicants are rejected and over 1500 applicants were approved for their home loan.



By Credit History meets guidelines - gdlin

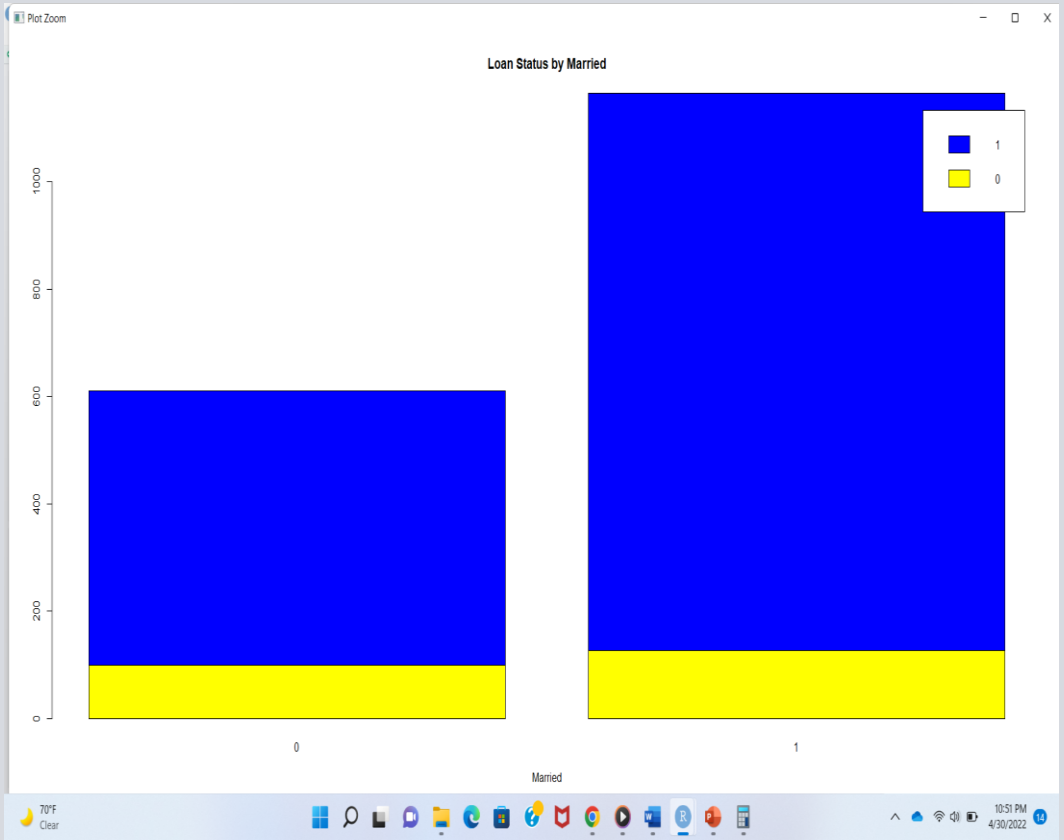
Loan Status by credit history meets guidelines – On x axis we have credit history, where 0 indicated it does not meet credit guidelines while 1 means they met credit guidelines and on legend 0 indicates rejected while 1 indicates approved.

The data shows there are more applicants who have more cleared the credit history guidelines. Surprisingly there are a few applicants who didn’t meet the credit guidelines but still got their loan approved. Even after meeting the credit history guidelines applicants were still rejected on some other basis.

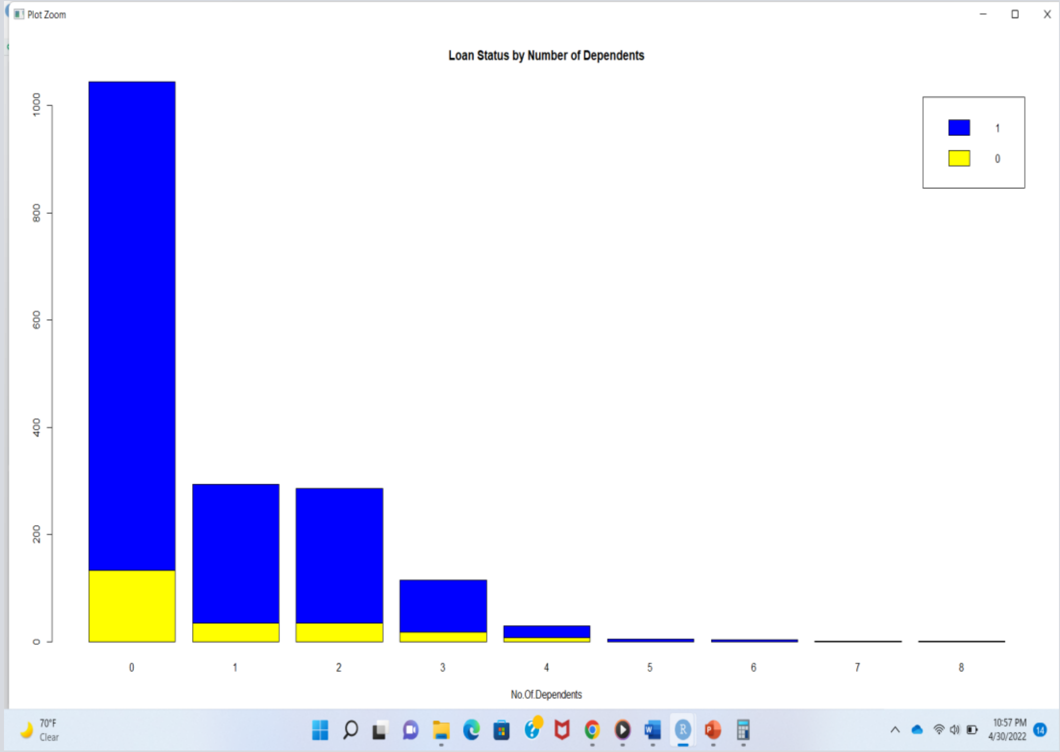


Marital status & Number of Dependents

Loan Status by Married/Unmarried – 0 indicates Unmarried, 1 indicates Married on x axis and on legend 0 indicates rejected while 1 indicates approved. There are a smaller number of Applicants who are single who have applied for loan than married applicants. Also, rejection rate of single applicants are more than married.

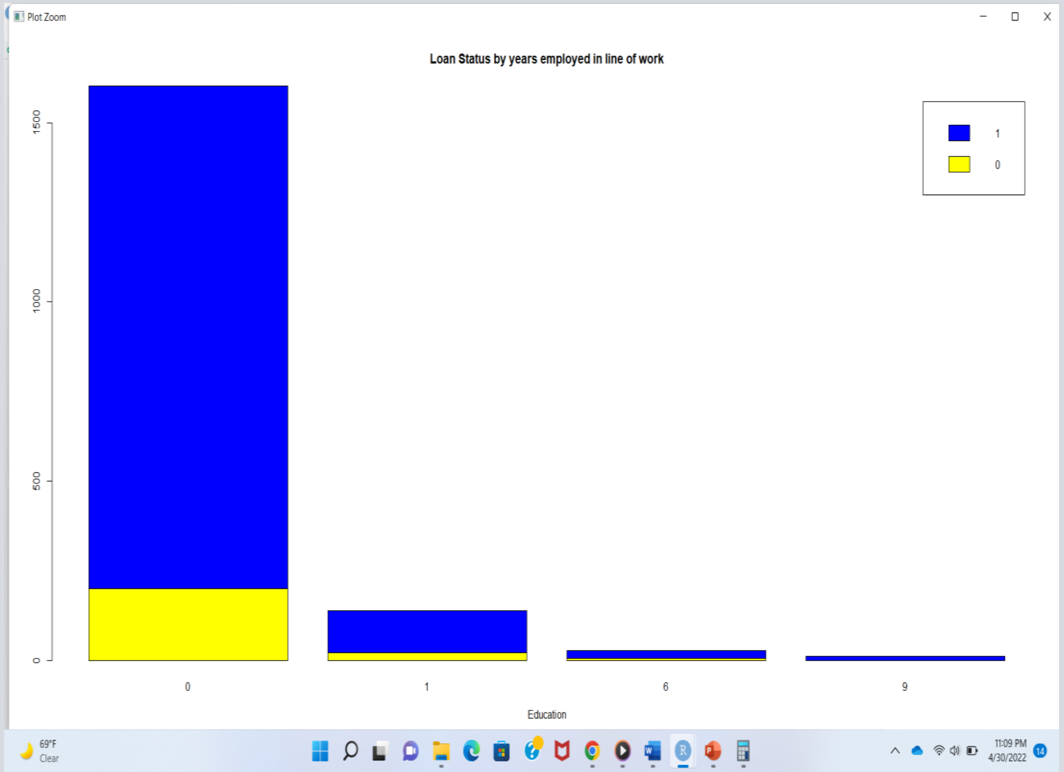


Loan Status by Number of Dependents– On x axis we have number of dependent ranging from 0-8 and on legend 0 indicates rejected while 1 indicates approved. The data shows there are more applicants who have applied for loan and have no dependents or up to 3 dependents where else people with more dependents are less but chances of getting them for loan approval is more.

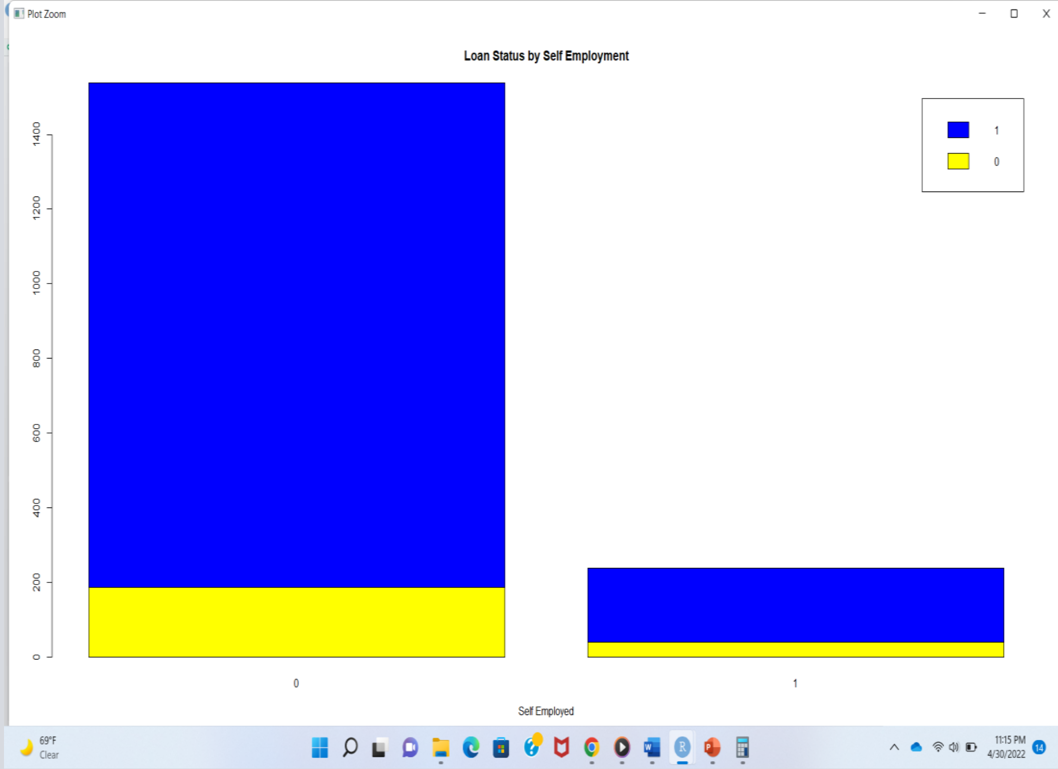


Employed vs Self-employed

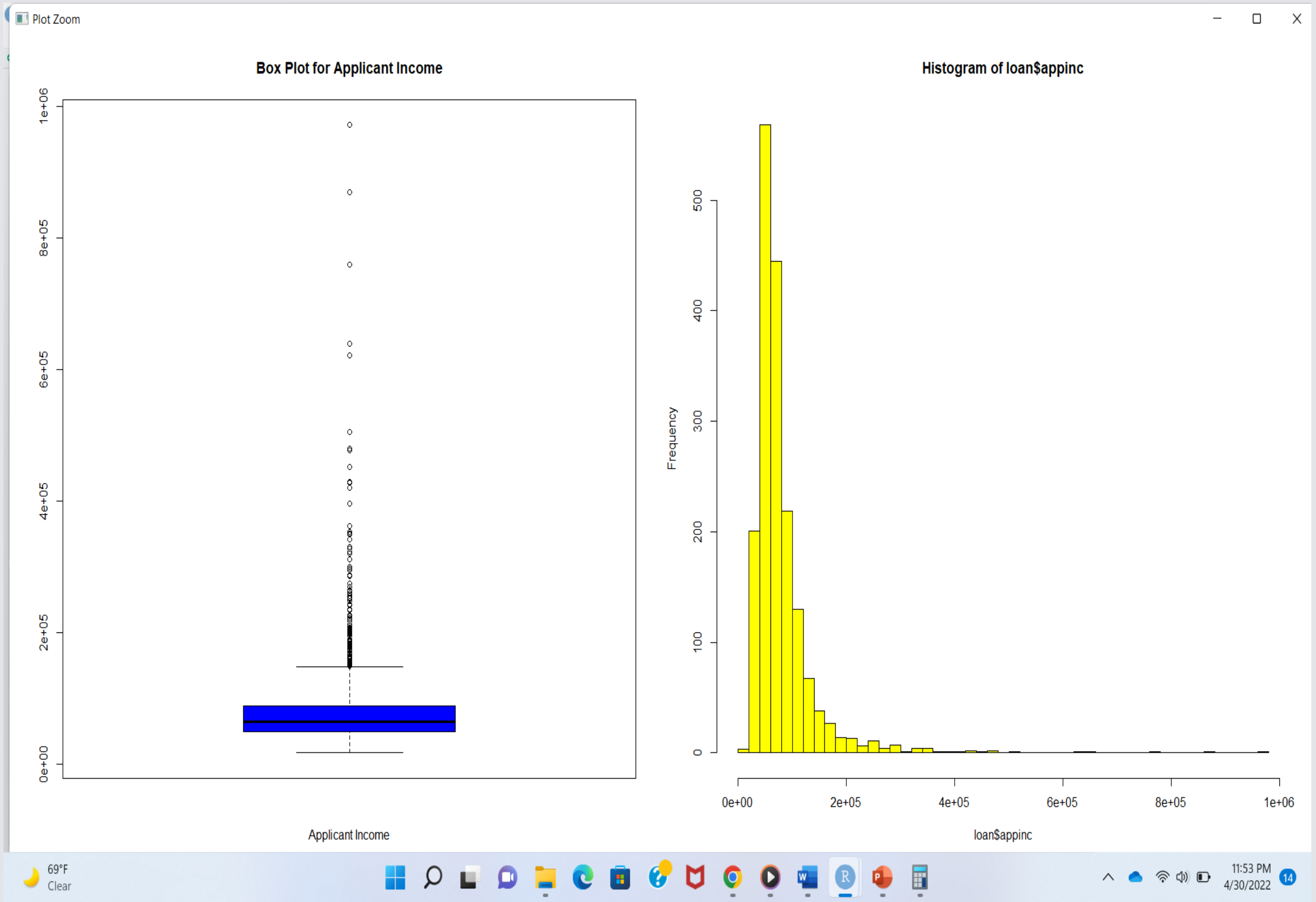
Loan Status by years employed in line of work – On x axis we have number of years employed in line of work ranging from 0-9 and on legend 0 indicates rejected while 1 indicates approved. The data shows there are more applicants who have applied for loan and have number of work experience from 0-1 years where else people with more work experience are less but chances of getting them for loan approval is more.



Loan Status by self employed – On x axis we have 0 indicating not self employed(working class) and 1 indicated self-employed and on legend 0 indicates rejected while 1 indicates approved. The data shows there are more applicants who have applied for loan and are from working class where else people who are self employed apply less for loan more. Approve rate for working class is more.



By Applicant Income – appinc



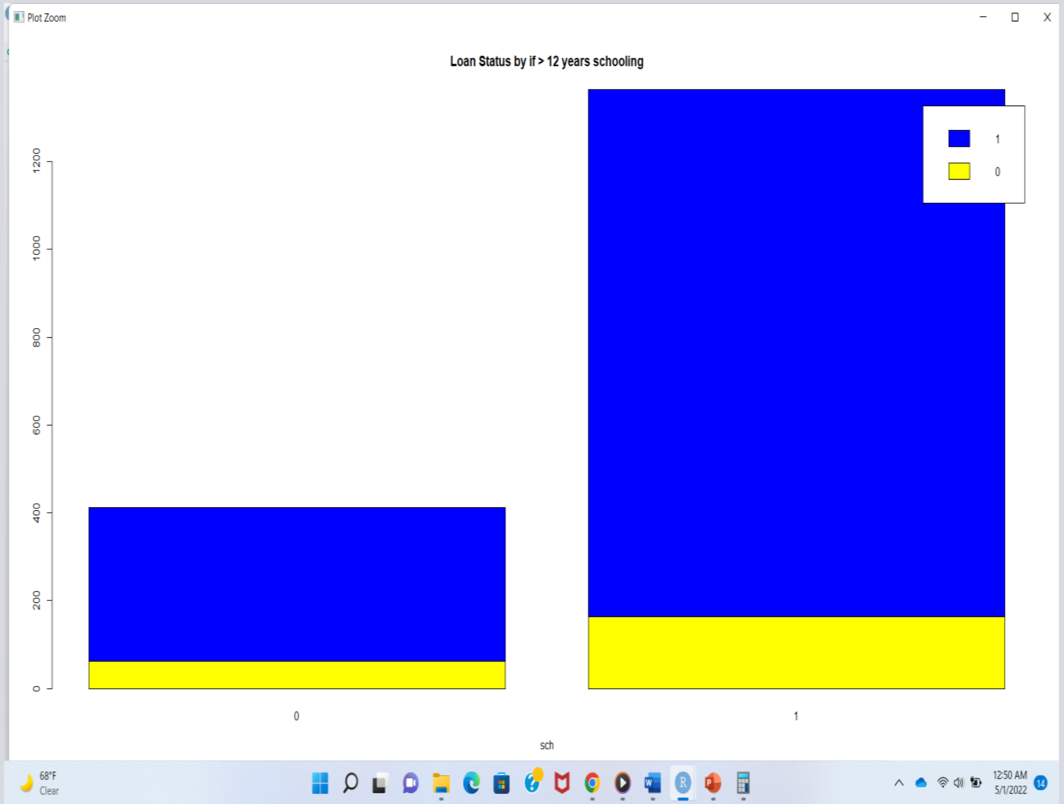
On the x axis we can see applicant income in 1000's$ and on y axis we can see number of applicants in histogram. And in box plot on y axis there is applicant income box plot

We can infer that people having income between 0e+00 to 2e+05 have applied more for loan than others. The graph is right skewed

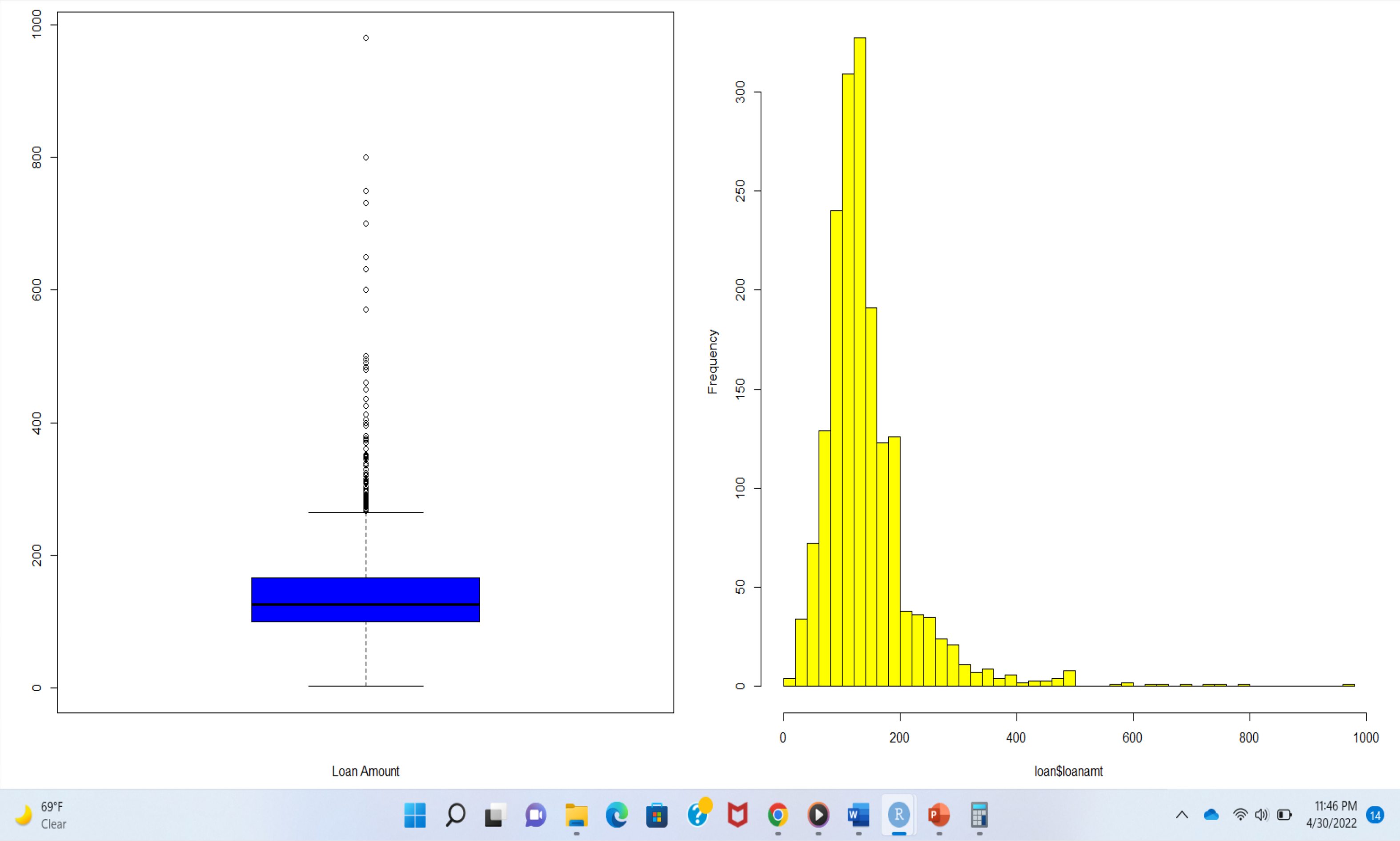
Evaluation by Gender & Education

Loan Status by Gender – 0 indicates female, 1 indicates male on x axis and on legend 0 indicates rejected while 1 indicates approved. More men have applied for home loan than females and more men have gotten approved for their loan than female.

Loan Status by Education – 0 indicates applicants < 12 years of schooling, 1 indicates applicants > 12 years of schooling on x axis and on legend 0 indicates rejected while 1 indicates approved. More educated applicants have applied for home loan than uneducated applicants and more educated class have gotten approved for their loan than uneducated category.



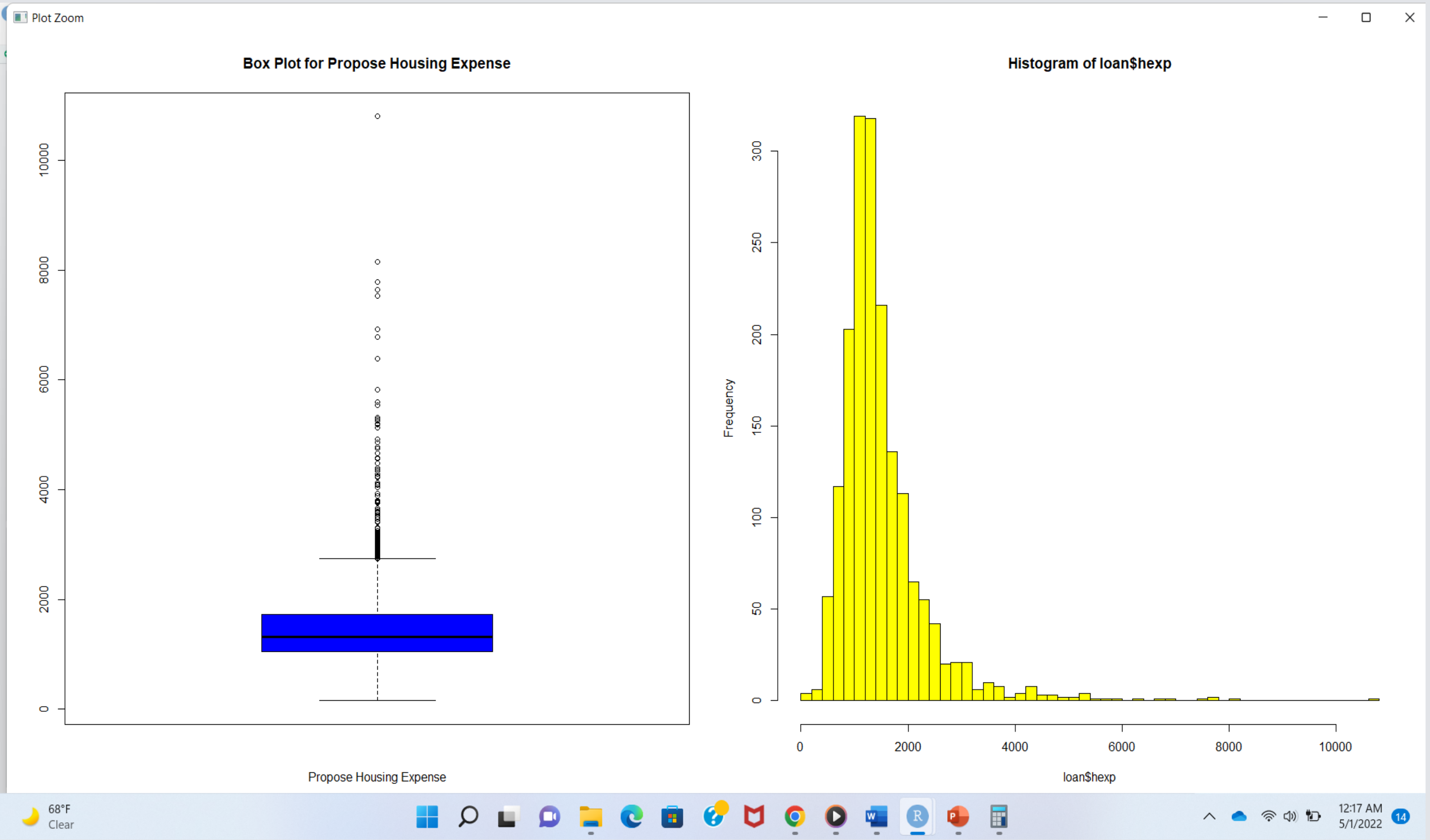
By Loan Amount – loanamt



On the x axis we can see number of applicants and on y axis we can see loan amount in thousands in histogram. And in box plot on y axis there is loan amount on box plot

We can see that there are more applicants who got their loan amount approved from 240,000 to 180,000 USD and the median is approximately 150,000 USD for the whole. The graph is right skewed

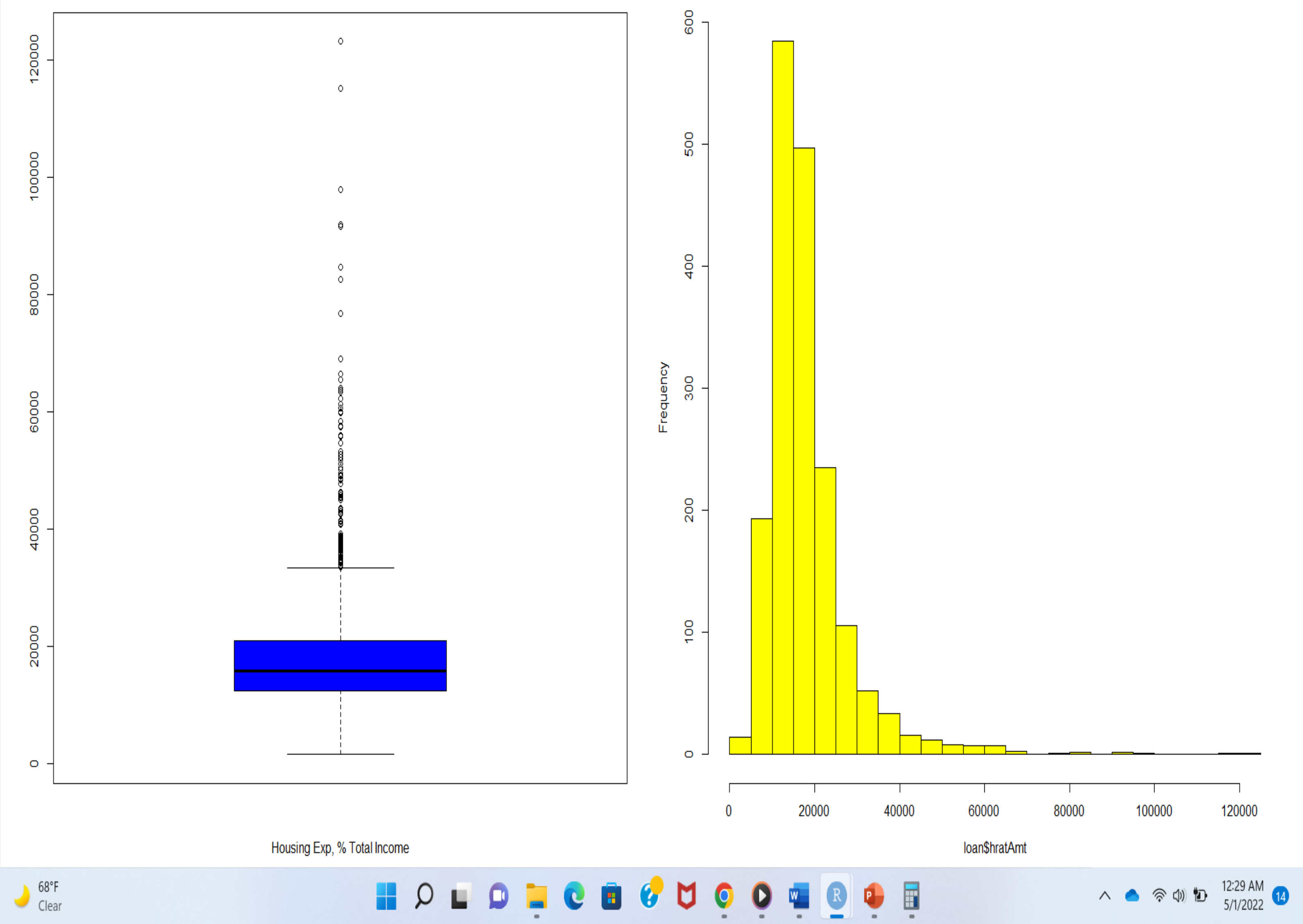
By Propose Housing Expense – hexp



On the x axis we can see proposed housing expense and on y axis we can see number of applicants in histogram. And in box plot on y axis there is proposed housing expense on box plot

We can see that there are more applicants who's housing expense is from 500 to 3000 USD and the median is approximately 1700 USD for the whole. It is a right skewed graph

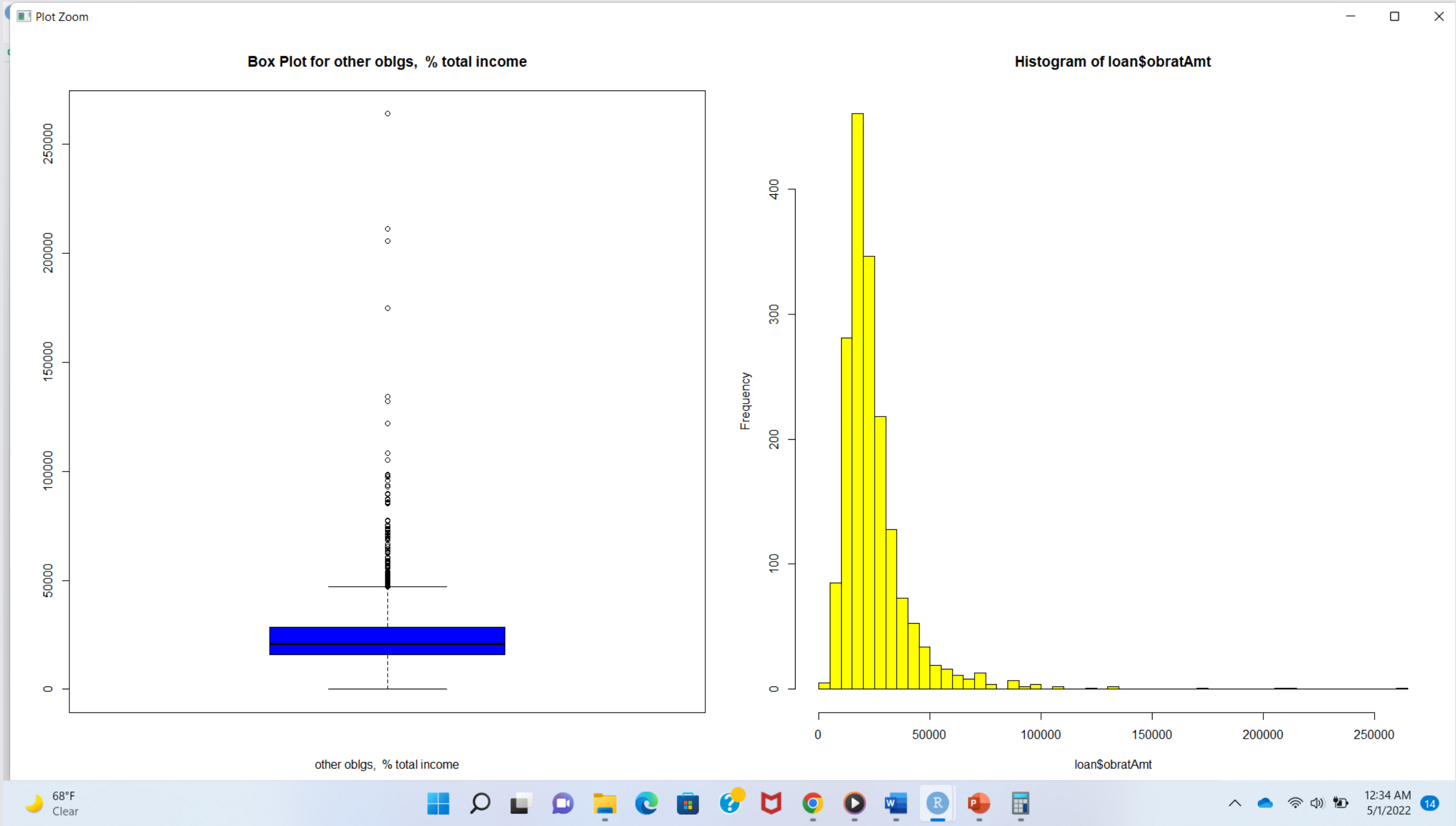
By Housing Exp, % Total Income – hratAmt



On the x axis we can see housing exp % total income and on y axis we can see number of applicants in histogram. And in box plot on y axis there is housing exp, % total income on box plot

We can see that there are more applicants who's housing expense % total income is from 5000 to 30000 USD and the median is approximately 17000 USD for the whole. It is a right skewed graph

By other obligations, % total income – obratAmt



On the x axis we can see other oblgs, % total income and on y axis we can see number of applicants in histogram. And in box plot on y axis there is other oblgs, % total income on box plot

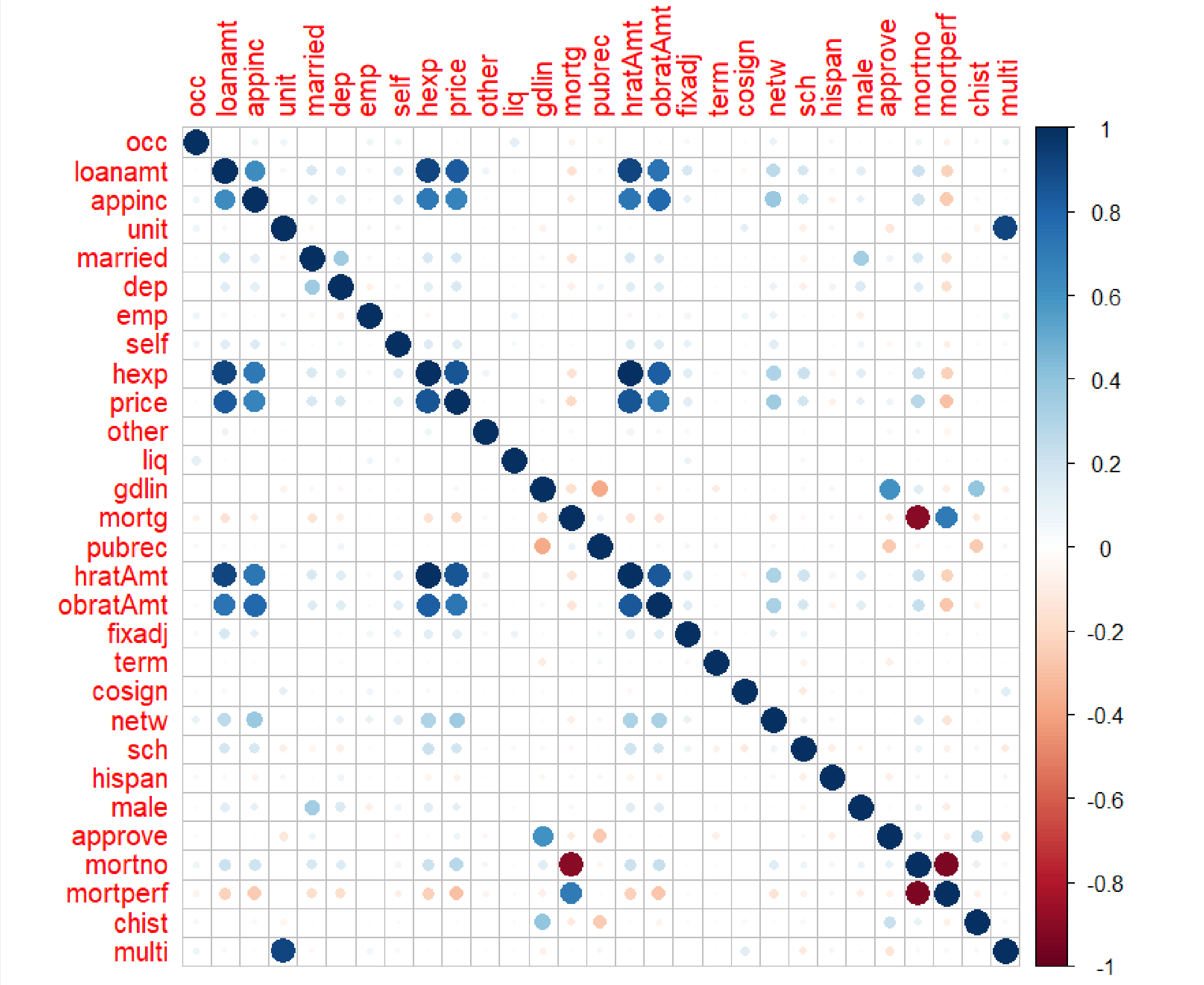
We can see that there are more applicants who's other oblgs, % total income is from 10000 to 25000 USD and the median is approximately 15000 USD for the whole. It is a right skewed graph

Descriptive Statistics

```
Console  Terminal ×  Jobs ×
R 4.1.2 · ~/
> summary(loan.df)
      occ             loanamt           appinc              unit          married       dep                emp              self            hexp      
 Min.   :1.000   Min.   :  2.0   Min.   : 17304   Min.   :1.000   0: 611   Min.   :0.0000   Min.   :0.0000   0:1538   Min.   :  154  
 1st Qu.:1.000   1st Qu.:100.0   1st Qu.: 49803   1st Qu.:1.000   1:1165   1st Qu.:0.0000   1st Qu.:0.0000   1: 238   1st Qu.: 1043  
 Median :1.000   Median :126.0   Median : 64392   Median :1.000            Median :0.0000   Median :0.0000            Median : 1312  
 Mean   :1.033   Mean   :143.2   Mean   : 80148   Mean   :1.126            Mean   :0.7776   Mean   :0.2162            Mean   : 1505  
 3rd Qu.:1.000   3rd Qu.:166.0   3rd Qu.: 89112   3rd Qu.:1.000            3rd Qu.:1.0000   3rd Qu.:0.0000            3rd Qu.: 1724  
 Max.   :3.000   Max.   :980.0   Max.   :972000   Max.   :4.000            Max.   :8.0000   Max.   :9.0000            Max.   :10798  
     price           other               liq          gdlin        mortg          pubrec      hratAmt           obratAmt        fixadj  
 Min.   :  25   Min.   :  0.000   Min.   :      0   0: 158   Min.   :1.000   0:1648   Min.   :  1557   Min.   :     0   0:1219  
 1st Qu.: 129   1st Qu.:  0.000   1st Qu.:     20   1:1618   1st Qu.:1.000   1: 128   1st Qu.: 12503   1st Qu.: 15877   1: 557  
 Median : 162   Median :  0.000   Median :     38            Median :2.000            Median : 15780   Median : 20672           
 Mean   : 196   Mean   :  1.937   Mean   :   4596            Mean   :1.707            Mean   : 18123   Mean   : 24701           
 3rd Qu.: 225   3rd Qu.:  0.000   3rd Qu.:     83            3rd Qu.:2.000            3rd Qu.: 20886   3rd Qu.: 28331           
 Max.   :1535   Max.   :410.000   Max.   :1000000            Max.   :4.000            Max.   :123127   Max.   :263993           
      term          cosign        netw               sch        hispan     male      approve    mortno    mortperf chist     multi   
 Min.   :     6   0:1727   Min.   :-7919.00   0: 412   0:1675   0: 338   0: 225   0:1185   0: 644   0: 291   0:1619  
 1st Qu.:   360   1:  49   1st Qu.:   42.38   1:1364   1: 101   1:1438   1:1551   1: 591   1:1132   1:1485   1: 157  
 Median :   360            Median :   94.00                                                                            
 Mean   :  1466            Mean   :  248.78                                                                            
 3rd Qu.:   360            3rd Qu.:  230.00                                                                            
 Max.   :999999            Max.   :23448.00                                                                            
> View(summary(loan.df))
```

Correlation Matrix

- Number of late mortgage payment is negatively correlated with number of mortgage history.
- Number of mortgage history is negatively correlated with credit history in mortgage housing exp % total income is positively correlated with loan amount , housing expense.
- Multi is positively correlated with Unit.



Target Variable

- Approve is the target variable / dependent variable in our dataset.
- Approve is the status of the customer’s loan.
- Approve is the Boolean value.
- Here we are trying to predict if the customer’s loan will be approved or not depending upon customer’s data that we feed in the model.
- If the approve = 1 then customer’s loan is approved, else approve = 0 then customer’s loan is rejected.

Goals and Objectives

- Predictive modeling is a form of data mining that analyzes historical data with the goal of identifying trends or patterns and then using those insights to predict future outcomes.
- Here we are trying to identify the trends or patterns of the loan approval based on the historical data.
- With the predictive models that we train, we are trying to predict the outcome whether the loan should be approved or not.
- Goal of the models should predict the outcome with the minimum error and should not be biased.

Build the Models

Here we are training 3 models to predict the outcome:

- Logistic Regression
- Decision Tree
- Random Forest

Logistic Regression

```
Call:
glm(formula = approve ~ ., family = "binomial", data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0330   0.1971   0.2666   0.3687   2.0767

Coefficients: (1 not defined because of singularities)
                Estimate    Std. Error z value              Pr(>|z|)
(Intercept)  1.0818559009 4.3873527731   0.247              0.805229
occ         -0.4327151506 0.5172664409  -0.837              0.402850
loanamt     -0.0082339661 0.0043815659  -1.879              0.060213 .
msa                    NA           NA      NA                    NA
appinc       0.0000042166 0.0000031225   1.350              0.176890
unit        -0.3838106891 0.4195308690  -0.915              0.360267
married1     1.0725418830 0.2804493812   3.824              0.000131 ***
dep         -0.1582880835 0.1153282312  -1.373              0.169908
emp         -0.0943827071 0.1061983767  -0.889              0.374143
self1       -0.6990873524 0.2892781372  -2.417              0.015664 *
hexp        -0.0012788456 0.0010741393  -1.191              0.233820
price        0.0064728444 0.0030222929   2.142              0.032218 *
other       -0.0077941317 0.0038284317  -2.036              0.041765 *
liq         -0.0000006861 0.0000011430  -0.600              0.548320
gdlin1       3.9087601917 0.3450003348  11.330 < 0.0000000000000002 ***
mortg       -0.1285211879 1.2384574249  -0.104              0.917348
pubrec1     -0.6014961800 0.3548246684  -1.695              0.090039 .
hratAmt      0.0001291841 0.0000878041   1.471              0.141216
obratAmt    -0.0000313003 0.0000138701  -2.257              0.024029 *
fixadj1      0.2953774580 0.2616971672   1.129              0.259025
term        -0.0017252326 0.0020447655  -0.844              0.398820
cosign1      0.6857667857 0.7722010117   0.888              0.374504
netw        -0.0000581148 0.0002615801  -0.222              0.824183
sch1         0.2257493127 0.2638142114   0.856              0.392156
hispan1     -0.7890130337 0.4171835451  -1.891              0.058586 .
male1       -0.3045755530 0.3253150395  -0.936              0.349146
mortno1     -0.6592253991 3.0246613008  -0.218              0.827468
mortperf1   -0.4766022932 1.8184632486  -0.262              0.793252
chist1      -0.1672762503 0.3266228963  -0.512              0.608554
multi1      -0.5709071441 0.7167310952  -0.797              0.425716
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 935.19  on 1242  degrees of freedom
Residual deviance: 579.81  on 1214  degrees of freedom
AIC: 637.81

Number of Fisher Scoring iterations: 11
```

```
Confusion Matrix and Statistics


       FALSE TRUE
  FALSE   34   10
  TRUE    36  453

               Accuracy : 0.9137
                 95% CI : (0.8866, 0.9361)
    No Information Rate : 0.8687
    P-Value [Acc > NIR] : 0.0007731

                  Kappa : 0.551

 Mcnemar's Test P-Value : 0.0002278

            Sensitivity : 0.48571
            Specificity : 0.97840
         Pos Pred Value : 0.77273
         Neg Pred Value : 0.92638
             Prevalence : 0.13133
         Detection Rate : 0.06379
   Detection Prevalence : 0.08255
      Balanced Accuracy : 0.73206

       'Positive' Class : FALSE
```

Even if the accuracy of the model is 91.37% but in the confusion matrix the percent of not approved correctly predicted as not approved is 34/70 = 48.6%. Which means our model is biased and cannot be implemented.

Logistic Regression with Reduced columns

```
Call:
glm(formula = approve ~ ., family = "binomial", data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9669   0.2342   0.2905   0.3991   2.0883

Coefficients:
               Estimate   Std. Error z value            Pr(>|z|)
(Intercept) -1.182479397  0.349803350  -3.380            0.000724 ***
loanamt     -0.007785049  0.003482710  -2.235            0.025395 *
married1     0.835781245  0.229752927   3.638            0.000275 ***
self1       -0.789310713  0.278522441  -2.834            0.004598 **
price        0.007490548  0.002713047   2.761            0.005764 **
gdlin1       3.676333115  0.282499968  13.014 < 0.0000000000000002 ***
other       -0.005444092  0.003828135  -1.422            0.154989
pubrec1     -0.635420322  0.339668322  -1.871            0.061386 .
obratAmt    -0.000011431  0.000008814  -1.297            0.194671
hispan1     -0.778427384  0.397743605  -1.957            0.050335 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 935.19  on 1242  degrees of freedom
Residual deviance: 609.10  on 1233  degrees of freedom
AIC: 629.1

Number of Fisher Scoring iterations: 6
```

Taking into consideration only the columns which are significant to the logistic regression will this have any impact on the confusion matrix and the accuracy of the model?

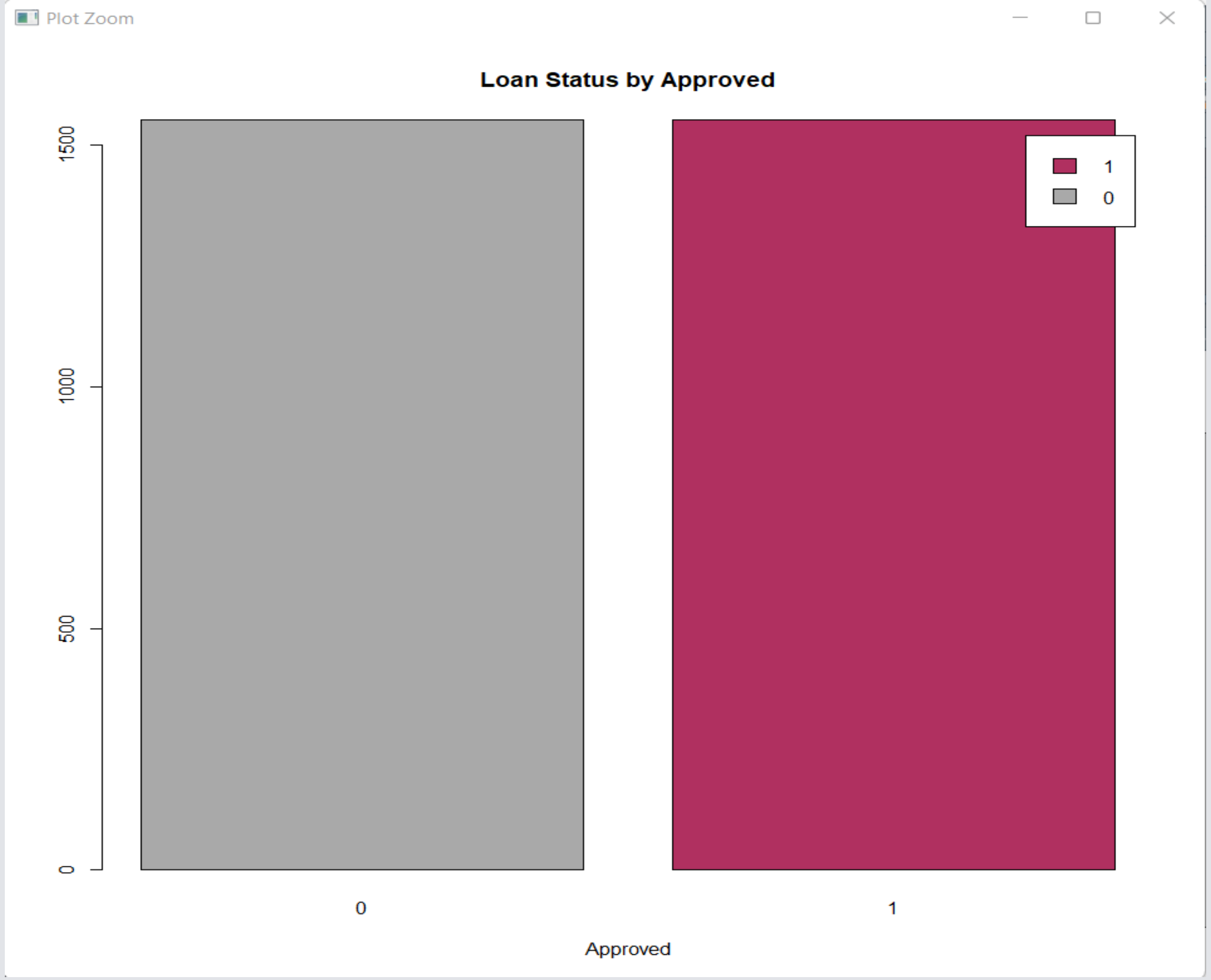
Logistic Regression with Reduced columns

```
Confusion Matrix and Statistics


       FALSE TRUE
  FALSE    36   10
  TRUE     34  453

               Accuracy : 0.9174
                 95% CI : (0.8908, 0.9394)
    No Information Rate : 0.8687
    P-Value [Acc > NIR] : 0.0002745

                  Kappa : 0.5766

 Mcnemar's Test P-Value : 0.0005256

            Sensitivity : 0.51429
            Specificity : 0.97840
         Pos Pred Value : 0.78261
         Neg Pred Value : 0.93018
             Prevalence : 0.13133
         Detection Rate : 0.06754
   Detection Prevalence : 0.08630
      Balanced Accuracy : 0.74634

       'Positive' Class : FALSE
```

There was no significant difference in the accuracy and the confusion matrix compared to previous model where all the columns were taken into consideration.
This is because of the data imbalance. In the dataset we have many rows with approve = 1 but very less data for approve = 0 hence our model is not able to accurately predict the loan denial.

Hence upsampling the data.

Upsampling of Dataset



Due to imbalance between the number of approve and rejects, we have upsampled the data.

Logistic Regression with Upsampling

```
Call:
glm(formula = approve ~ ., family = "binomial", data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1877  -0.4728   0.2227   0.7931   3.1730

Coefficients: (1 not defined because of singularities)
                Estimate    Std. Error z value              Pr(>|z|)
(Intercept) -7.93535808205 2.30249293790  -3.446             0.000568 ***
occ         -0.77680777330 0.24683597081  -3.147             0.001649 **
loanamt     -0.00629611050 0.00214566030  -2.934             0.003343 **
msa                     NA            NA      NA                   NA
appinc       0.00000657072 0.00000179623   3.658             0.000254 ***
unit        -0.21154193973 0.26877975580  -0.787             0.431255
married1     0.46147754703 0.12969315243   3.558             0.000373 ***
dep         -0.01415657311 0.05694402698  -0.249             0.803666
emp         -0.14626630301 0.05562261109  -2.630             0.008548 **
self1       -0.69505560041 0.15227138432  -4.565            0.0000050 ***
hexp        -0.00139097481 0.00055091589  -2.525             0.011575 *
price        0.00451462957 0.00128573637   3.511             0.000446 ***
other       -0.00687674415 0.00303457050  -2.266             0.023443 *
liq          0.00000003023 0.00000073502   0.041             0.967192
gdlin1       3.87334782471 0.25239114544  15.347 < 0.0000000000000002 ***
mortg        1.70305907769 0.66100276409   2.576             0.009981 **
pubrec1     -0.51528719512 0.22584043245  -2.282             0.022510 *
hratAmt      0.00012875643 0.00004435095   2.903             0.003695 **
obratAmt    -0.00002594397 0.00000733390  -3.538             0.000404 ***
fixadj1      0.51195714215 0.12905182847   3.967            0.0000728 ***
term        -0.00113574666 0.00094235774  -1.205             0.228119
cosign1      1.42767095489 0.41940177509   3.404             0.000664 ***
netw         0.00001047398 0.00008965853   0.117             0.907002
sch1        -0.00356022344 0.13606880403  -0.026             0.979126
hispan1     -0.89981748159 0.20558415405  -4.377            0.0000120 ***
male1       -0.03108073077 0.15061113764  -0.206             0.836507
mortno1      4.40439718551 1.55086826746   2.840             0.004512 **
mortperf1    2.75583224596 0.90783902779   3.036             0.002401 **
chist1      -0.01900472838 0.17201277276  -0.110             0.912025
multi1      -0.50894528522 0.42647018054  -1.193             0.232717
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3009.5  on 2170  degrees of freedom
Residual deviance: 2004.6  on 2142  degrees of freedom
AIC: 2062.6

Number of Fisher Scoring iterations: 13
```

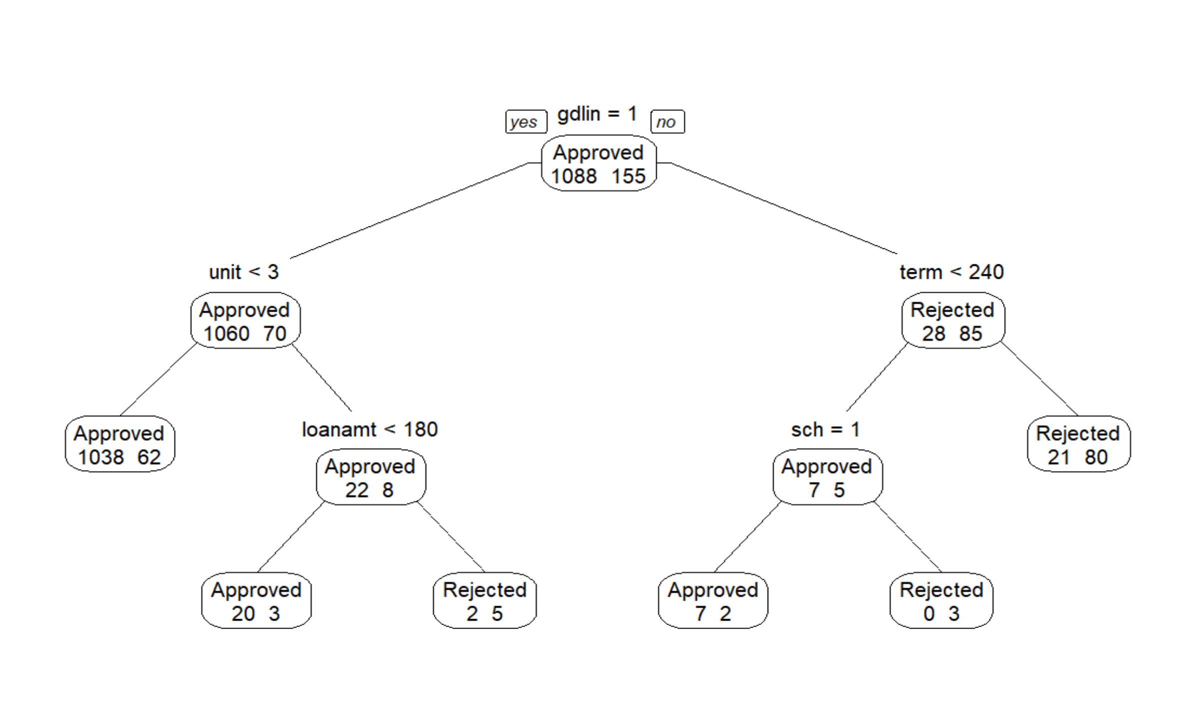
```
Confusion Matrix and Statistics


       FALSE TRUE
 FALSE   303   51
 TRUE    170  407

               Accuracy : 0.7626
                 95% CI : (0.7339, 0.7896)
    No Information Rate : 0.5081
    P-Value [Acc > NIR] : < 0.0000000000000002

                  Kappa : 0.5271

 Mcnemar's Test P-Value : 0.000000000000002062

            Sensitivity : 0.6406
            Specificity : 0.8886
         Pos Pred Value : 0.8559
         Neg Pred Value : 0.7054
             Prevalence : 0.5081
         Detection Rate : 0.3255
   Detection Prevalence : 0.3802
      Balanced Accuracy : 0.7646

       'Positive' Class : FALSE
```

Upsampling the data reduced the accuracy of the model but has increased the percentage of not approved correctly predicted as not approved to 303/473 = 64%
But still this model cannot be implemented due to overall accuracy of the model being only 76%.
Hence, not selecting logistic regression for prediction.

Decision Tree

Here we have the accuracy of 89.31% but our model predicts the rejected correctly as rejected with accuracy of only 31/70 = 44.28%. Hence this model will most of the time approve the loan and this cannot be implemented.

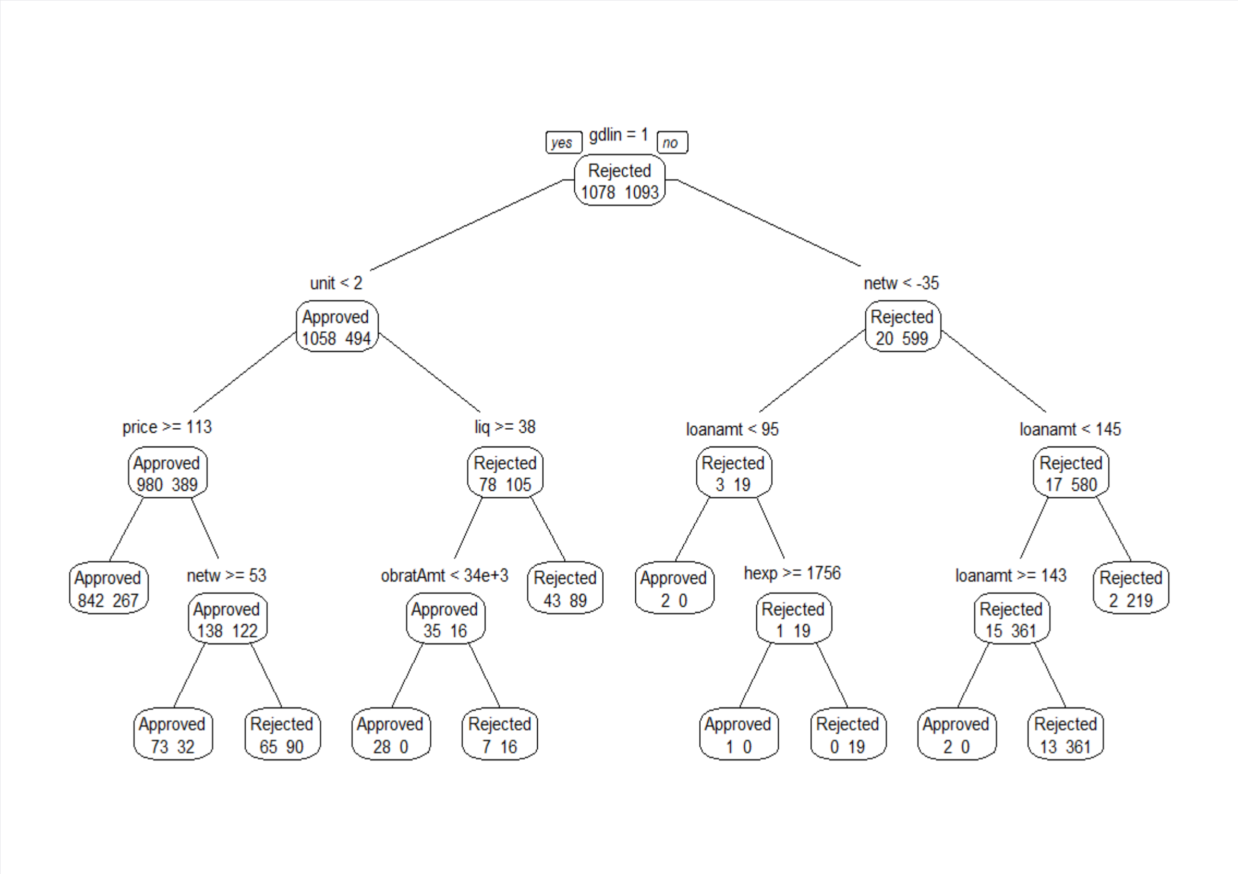
```
> confusionMatrix(ct.pred, as.factor(valid.df$approve))
Confusion Matrix and Statistics

          Reference
Prediction Approved Rejected
  Approved      445       39
  Rejected       18       31

               Accuracy : 0.8931
                 95% CI : (0.8637, 0.918)
    No Information Rate : 0.8687
    P-Value [Acc > NIR] : 0.051492

                  Kappa : 0.4629

 Mcnemar's Test P-Value : 0.008071

            Sensitivity : 0.9611
            Specificity : 0.4429
         Pos Pred Value : 0.9194
         Neg Pred Value : 0.6327
             Prevalence : 0.8687
         Detection Rate : 0.8349
   Detection Prevalence : 0.9081
      Balanced Accuracy : 0.7020

       'Positive' Class : Approved
```



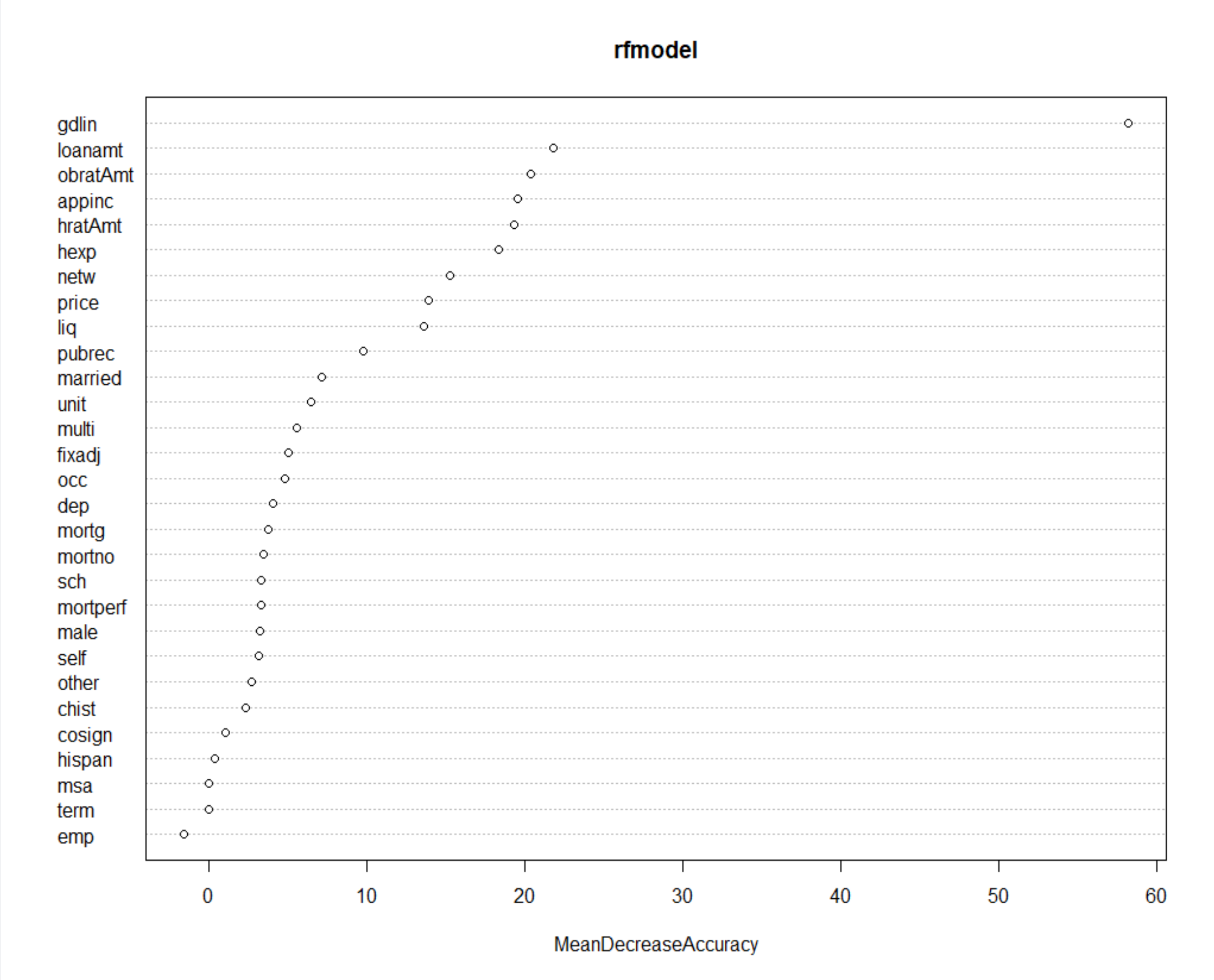
Decision Tree with Upsampling

Even upsampling the dataset only improves the accuracy of correctly predicting rejected as rejected from 45% to 336/458 = 73%. And overall accuracy of the model is only 80%

```
> confusionMatrix(ct.pred, as.factor(valid.df$approve))
Confusion Matrix and Statistics

          Reference
Prediction Approved Rejected
  Approved      409      122
  Rejected       64      336

               Accuracy : 0.8002
                 95% CI : (0.7731, 0.8255)
    No Information Rate : 0.5081
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5995

 Mcnemar's Test P-Value : 2.922e-05

            Sensitivity : 0.8647
            Specificity : 0.7336
         Pos Pred Value : 0.7702
         Neg Pred Value : 0.8400
             Prevalence : 0.5081
         Detection Rate : 0.4393
   Detection Prevalence : 0.5704
      Balanced Accuracy : 0.7992

       'Positive' Class : Approved
```



Random Forest



According to random forest the most important variable for accurately predicting is gdlin (credit history). According to the model the decision of the loan approval is mostly dependent on the gdlin.

So, let's explore our dataset and see if it's true or not.

Check is the dataset with approve =1 and gdlin = 1.

Data	
check	1514 obs. of 30 variables
loan	1776 obs. of 30 variables

There is total 1776 observations in total and out of which 1514 observations have approve =1 and gdlin = 1. Which means that the decision of the loan approval is mostly dependent on the gdlin.

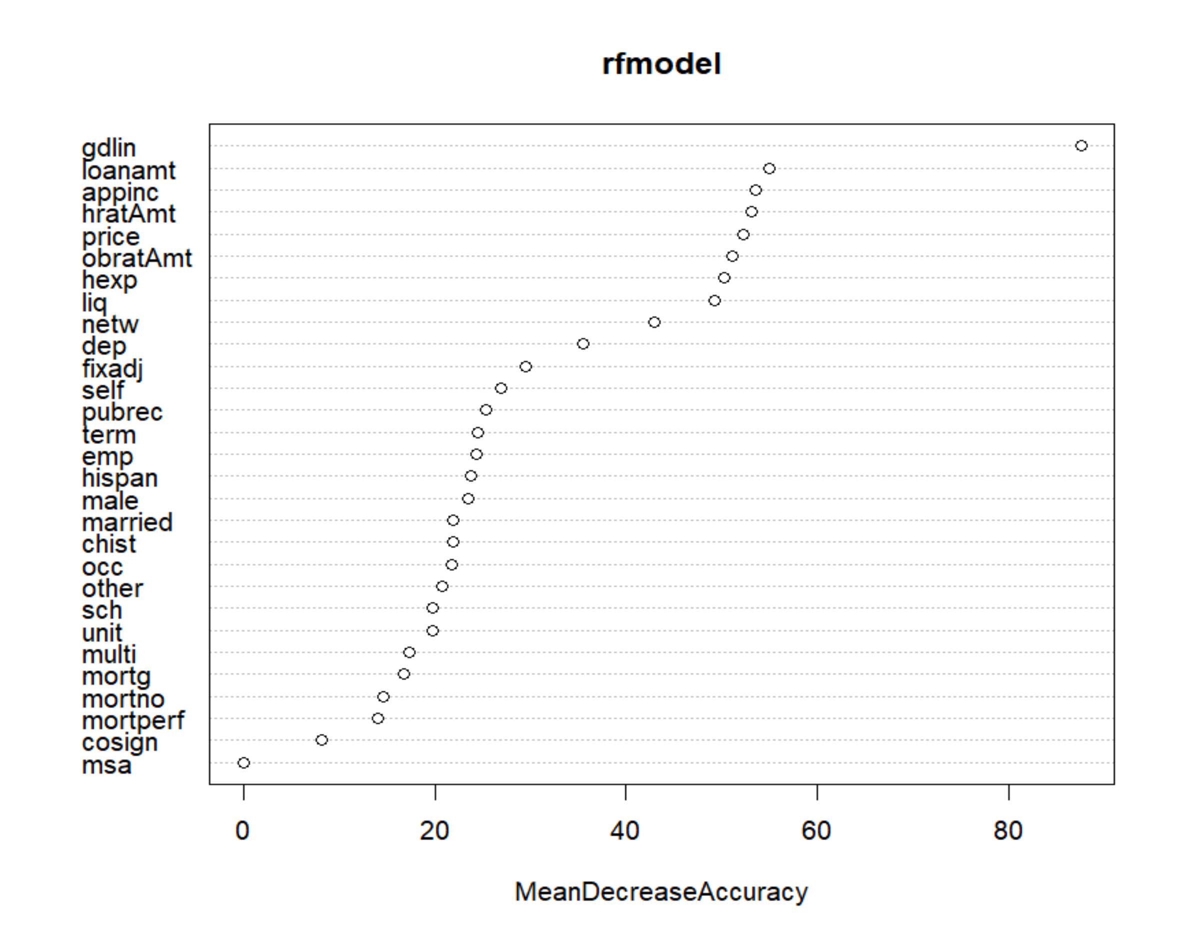
[redacted]

```
> confusionMatrix(rf.pred, as.factor(valid.df$approve))
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  31   9
         1  39 454

               Accuracy : 0.9099
                 95% CI : (0.8824, 0.9329)
    No Information Rate : 0.8687
    P-Value [Acc > NIR] : 0.001991

                  Kappa : 0.5176

 Mcnemar's Test P-Value : 0.00002842

            Sensitivity : 0.44286
            Specificity : 0.98056
         Pos Pred Value : 0.77500
         Neg Pred Value : 0.92089
             Prevalence : 0.13133
         Detection Rate : 0.05816
   Detection Prevalence : 0.07505
      Balanced Accuracy : 0.71171

       'Positive' Class : 0
```

The accuracy of the model is 90.99% but similar to logistic regression this model is also not able to accurately predict not approved as not approved (31/70 = 45%)

But still this model is not efficient and cannot be used in the real world.

Random Forest using Upsampling



Upsampling the data helped us to understand that gdlin is dominant variable but there are other variables which are dominant enough to impact the model like loanamt, appinc, hrat, price, obratAmt, hexp etc.

[redacted]

```
> confusionMatrix(rf.pred, as.factor(valid.df$approve))
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 473  13
         1   0 445

               Accuracy : 0.986
                 95% CI : (0.9762, 0.9925)
    No Information Rate : 0.5081
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.9721

 Mcnemar's Test P-Value : 0.0008741

            Sensitivity : 1.0000
            Specificity : 0.9716
         Pos Pred Value : 0.9733
         Neg Pred Value : 1.0000
             Prevalence : 0.5081
         Detection Rate : 0.5081
   Detection Prevalence : 0.5220
      Balanced Accuracy : 0.9858

       'Positive' Class : 0
```

Random Forest Model with upsampling gives us the accuracy of 98.6% which can be implemented in real world. Also, model can predict not approved as not approved correctly with 100% accuracy hence this model is not biased.

Model Evaluation

Model	Accurately Predicted Rejected as Rejected	Overall Accuracy
Logistic Regression	48.6%	91.37%
Logistic Regression with selected columns	51.4%	91.74%
Logistic Regression with Upsampling	64%	76.26%
Decision Tree	44.28%	89.31%
Decision Tree with Upsampling	73%	80%
Random Forest	45%	90.99%
Random Forest with Upsampling	100%	98.6%

Recommendations

Based on our final model, the three ways JP Morgan Chase Bank will able to use the model are:

- The bank does not have to go through the supporting documents of all the customers. It will only have to go through the supporting documents of the customers who are approved by the algorithm. This saves the manpower used by the bank by decreasing the burden of going through the supporting documentation of all the customers.
- Customer satisfaction may increase as the bank response time decreases.
- The chances of the bank losing the money, through scams, decreases.

Thankyou