**SOLUTIONS**

**1.** Total rate of Default=0.05.

**data** combine;

set loandetails36 loandetails60;

**run**;

**proc** **sort** data =combine;

by loan\_status;

**run**;

**proc** **freq** data = combine;

table loan\_status;

**run**;

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **loan\_status** | | | | |
| **loan\_status** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Charged Off** | 1517 | 11.58 | 1517 | 11.58 |
| **Current** | 2007 | 15.32 | 3524 | 26.90 |
| **Default** | 6 | 0.05 | 3530 | 26.94 |
| **Fully Paid** | 9425 | 71.94 | 12955 | 98.88 |
| **In Grace Period** | 42 | 0.32 | 12997 | 99.20 |
| **Late (16-30 days)** | 10 | 0.08 | 13007 | 99.27 |
| **Late (31-120 days)** | 95 | 0.73 | 13102 | 100.00 |

**2.** Difference In default rate between 3 years and 5 years is 5%.

**data** loandetails36;

set loandetails36;

loan\_term= 3;

**run**;

**data** loandetails60;

set loandetails60;

loan\_term= 5 ;

**run**;

**data** combine;

set loandetails36 loandetails60;

**run**;

**proc** **freq** data =combine;

table loan\_term\*loan\_status/norow nocol;

**run**;

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table of loan\_term by loan\_status** | | | | | | | | |
| **loan\_term** | **loan\_status (loan\_status)** | | | | | | | **Total** |
| **Frequency Percent** | **Charged Off** | **Current** | **Default** | **Fully Paid** | **In Grace Period** | **Late (16-30 days)** | **Late (31-120 days)** |
| **3** | 1344 10.26 | 1259 9.61 | 6 0.05 | 9116 69.58 | 24 0.18 | 9 0.07 | 67 0.51 | 11825 90.25 |
| **5** | 173 1.32 | 748 5.71 | 0 0.00 | 309 2.36 | 18 0.14 | 1 0.01 | 28 0.21 | 1277 9.75 |
| **Total** | **1517 11.58** | **2007 15.32** | **6 0.05** | **9425 71.94** | **42 0.32** | **10 0.08** | **95 0.73** | **13102 100.00** |

**3.** Most common reason for Loan is Debt consolidation.

**proc** **freq** data=demo;

table purpose;

**run**;

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **purpose** | | | | |
| **purpose** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **car** | 396 | 3.02 | 396 | 3.02 |
| **credit\_card** | 1705 | 13.01 | 2101 | 16.04 |
| **debt\_consolidation** | 5637 | 43.02 | 7738 | 59.06 |
| **educational** | 314 | 2.40 | 8052 | 61.46 |
| **home\_improvement** | 897 | 6.85 | 8949 | 68.30 |
| **house** | 132 | 1.01 | 9081 | 69.31 |
| **major\_purchase** | 797 | 6.08 | 9878 | 75.39 |
| **medical** | 207 | 1.58 | 10085 | 76.97 |
| **moving** | 187 | 1.43 | 10272 | 78.40 |
| **other** | 1731 | 13.21 | 12003 | 91.61 |
| **renewable\_energy** | 24 | 0.18 | 12027 | 91.80 |
| **small\_business** | 658 | 5.02 | 12685 | 96.82 |
| **vacation** | 92 | 0.70 | 12777 | 97.52 |
| **wedding** | 325 | 2.48 | 13102 | 100.00 |

**4.** Average ratio of loan requested to Income is 0.178.

**proc** **sort** data =demo;

by id;

**run**;

**proc** **sort** data=combine;

by id;

**run**;

**data** final;

merge combine demo;

by id;

**run**;

**data** final;

set final;

ratio\_loan\_in = loan\_amnt /annual\_inc;

format ratio\_loan\_in 8.3;

**run**;

**proc** **means** mean data=final;

**run**;

|  |  |  |
| --- | --- | --- |
| **Variable** | **Label** | **Mean** |
| **id** | **id** | 451217.20 |
| **member\_id** | **member\_id** | 550598.05 |
| **loan\_amnt** | **loan\_amnt** | 10106.81 |
| **funded\_amnt** | **funded\_amnt** | 9865.67 |
| **funded\_amnt\_inv** | **funded\_amnt\_inv** | 9186.14 |
| **int\_rate** | **int\_rate** | 0.1193753 |
| **installment** | **installment** | 317.5810968 |
| **is\_inc\_v** | **is\_inc\_v** | 0 |
| **accept\_d** | **accept\_d** | 16OCT2009 |
| **issue\_d** | **issue\_d** | 25OCT2009 |
| **loan\_term** |  | 3.1949321 |
| **annual\_inc** | **annual\_inc** | 68391.51 |
| **ratio\_loan\_in** |  | 0.178 |

**5.** Average grant rate=1.028.

**data** final1;

set final;

grant\_rate = loan\_amnt / funded\_amnt;

format grant\_rate 8.3;

**run**;

**proc** **means** mean data=final1;

**run**;

|  |  |  |
| --- | --- | --- |
| **Variable** | **Label** | **Mean** |
| **id** | **id** | 451217.20 |
| **member\_id** | **member\_id** | 550598.05 |
| **loan\_amnt** | **loan\_amnt** | 10106.81 |
| **funded\_amnt** | **funded\_amnt** | 9865.67 |
| **funded\_amnt\_inv** | **funded\_amnt\_inv** | 9186.14 |
| **int\_rate** | **int\_rate** | 0.1193753 |
| **installment** | **installment** | 317.5810968 |
| **is\_inc\_v** | **is\_inc\_v** | 0 |
| **accept\_d** | **accept\_d** | 16OCT2009 |
| **issue\_d** | **issue\_d** | 25OCT2009 |
| **loan\_term** |  | 3.1949321 |
| **annual\_inc** | **annual\_inc** | 68391.51 |
| **ratio\_loan\_in** |  | 0.178 |
| **grant\_rate** |  | 1.028 |

**6.** Average grant rate differ by Income level.

**proc** **format** ;

value looky

0-<100000 = 1

100000-<200000=2

200000-<300000=3

300000-<400000=4

400000-<500000=5

500000 -high =6;

**run**;

**data** final2;

set final1;

format annual\_inc looky.;

**run**;

**proc** **means** noprint data= final2;

class annual\_inc;

output out = final3

mean(grant\_rate)=avg\_grant\_rate;

**run**;

**proc** **print** data=final3;

**run**;

INCOME GRANT\_RATE

1 1 11188 1.026

2 1 1661 1.036

3 1 170 1.050

4 1 37 1.033

5 1 8 1.428

6 1 38 1.036

AVG 0 13102 1.028

**7.** YES CA(CALIFORNIA) IS OVERPRESENTED IN DATE SET AS ITS POP\_SHARE IS 12% BUT LOAN\_SHARE IS 18%.

**proc** **sort** data=final2(rename =(addr\_state=states));

by states;

**run**;

**proc** **sort** data= population;

by states;

**run**;

**proc** **means** noprint data=population;/\*finding sum of total population of all states of US in different dataset and after that merged with population data set\*/

var \_2013;

output out=sum\_pop

sum(\_2013)=tot\_pop;

**run**;

**data** sum\_pop;

set sum\_pop;

common=1;

**run**;

**data** population;

set population;

common=1;

**run**;

**data** pop\_share(drop= \_type\_ \_freq\_ );

merge population sum\_pop;

by common;

ratio\_pop=(\_2013/tot\_pop)\*100;

**run**;

**proc** **means** noprint data=final2;

by states;

output out=loan sum(loan\_amnt)= ;

**run**;

**proc** **means** noprint data=final2;/\* finding sum of total loan amount of all states of US in different dataset and after that merging with loan data set\*/

output out=sum\_loan

sum(loan\_amnt)=tot\_loan

run;

**data** loan;

set loan;

common=1;

**run**;

**data** sum\_loan;

set sum\_loan;

common=1;

**run**;

**data** loan\_share(drop=\_freq\_ \_type\_);

merge loan sum\_loan;

by common;

loan\_percentage=(loan\_amnt/tot\_loan)\*100;

**run**;

**proc** **sort** data= pop\_share;

by states;

**run**;

**proc** **sort** data=loan\_share;

by states;

**run**;

**data** loan\_population\_comp(keep= states loan\_percentage ratio\_pop);/\*merged pop\_share and loan\_share having only three variable.\*/

merge pop\_share loan\_share;

by states;

**run**;

**proc** **sort** data= loan\_population\_comp;

by decending ratio\_pop;

**run**;

OUTPUT:

STATE POP\_SHARE LOAN\_SHARE

CA 12.12560079 18.678075365

TX 8.3662702472 7.0898397271

NY 6.2161766266 9.9799179765

FL 6.1850921485 6.9274202029

IL 4.0749635626 3.9936738889

PA 4.0406946232 3.9060356893

OH 3.6601557886 2.817864524

GA 3.1607894527 3.7510357714

MI 3.1302496891 2.0204362011

NC 3.1152045575 0.2699188582

NJ 2.8150987516 5.0643060865

VA 2.6129868525 4.0134406263

WA 2.2052420216 2.0174910139

MA 2.1171190902 4.0098346598

AZ 2.096178261 2.1004282416

IN 2.0785519033 0.065115069

TN 2.0548514399 0.1013257685

MO 1.9119328117 1.4126326255

MD 1.8754423098 2.8673474454

WI 1.8165735901 1.0859056366

MN 1.7146110482 1.556437056

CO 1.6665252739 2.3364774466

AL 1.5290354449 1.0031005647

SC 1.5104091785 1.1950285995

LA 1.463159772 0.8745506938

KY 1.3903492683 0.5994211197

OR 1.2431845865 1.0925700667

OK 1.2180375609 0.675354088

CT 1.1375362056 1.8856561264

IA 0.977581169 0.0426296973

MS 0.9461987111 0.1005328335

AR 0.9361287662 0.4532945223

UT 0.9176233365 0.6507731022

KS 0.9154359372 0.3488158931

NV 0.8825945804 1.0106712063

NM 0.6596320053 0.5968346411

NE 0.591061545 0.023939086

WV 0.5865659096 0.3180047036

ID 0.5099616995 0.0451217788

HI 0.4441398021 0.2727130102

ME 0.4201774201 0.0069476212

NH 0.4186454498 0.4207653069

RI 0.3326210299 0.5026830467

MT 0.3211238188 0.1505443782

DE 0.2928391484 0.3065260252

SD 0.2672571736 0.1170900719

AK 0.2325418973 0.1816765176

ND 0.2288285378 .

DC 0.2044890944 0.6734661474

VT 0.1982198151 0.1043275939

WY 0.1843102963 0.2636886544

XX . 0.0183130232

**8.** Impact of income verification status on default rate.

**proc** **freq** data=final2;

table is\_inc\_v\*loan\_status/out=default;

**run**;

OUTPUT: (1=INCOME VERFIED ,0=INCOME NOT VERFIED)

is\_inc\_v loan\_status COUNT PERCENT

1 Charged Off 404 3.0834987025

1 Current 666 5.0831934056

1 Default 2 0.0152648451

1 Fully Paid 2099 16.020454892

1 In Grace Period 16 0.1221187605

1 Late (16-30 days) 4 0.0305296901

1 Late (31-120 days) 22 0.1679132957

0 Charged Off 1113 8.4948862769

0 Current 1341 10.235078614

0 Default 4 0.0305296901

0 Fully Paid 7326 55.915127461

0 In Grace Period 26 0.1984429858

0 Late (16-30 days) 6 0.0457945352

0 Late (31-120 days) 73 0.5571668448

FOUR LOANS ARE DEFAULT WHEN INCOME NOT VERFIED AND THERE IS NO ANY DEFAULT LOAN WHEN INCOME IS VERIFIED.SO INCOME VERFICATION IS IMPORTANT AT THE TIME OF SANSATIONING THE LOAN.

**09.** AVERAGE INTEREST RATE BY GRADE AND SUB GRADE.

**proc** **means** mean data =final2 ;

class grade ;

var int\_rate;

**run**;

**proc** **means** mean data =final2 ;

class grade sub\_grade ;

var int\_rate;

**run**;

OUTPUT:

|  |  |  |
| --- | --- | --- |
| **Analysis Variable: int\_rate int\_rate** | | |
| **grade** | **N Obs** | **Mean** |
| A | 2985 | 0.0805357 |
| B | 4048 | 0.1115637 |
| C | 3183 | 0.1316957 |
| D | 1896 | 0.1491499 |
| E | 722 | 0.1660172 |
| F | 209 | 0.1829230 |
| G | 59 | 0.2029797 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Analysis Variable: int\_rate int\_rate** | | | |
| **grade** | **sub\_grade** | **N Obs** | **Mean** |
| A | A1 | 104 | 0.0687029 |
|  | A2 | 316 | 0.0720778 |
|  | A3 | 625 | 0.0756749 |
|  | A4 | 822 | 0.0823568 |
|  | A5 | 1118 | 0.0854055 |
| B | B1 | 655 | 0.1039140 |
|  | B2 | 736 | 0.1070061 |
|  | B3 | 816 | 0.1110619 |
|  | B4 | 892 | 0.1145900 |
|  | B5 | 949 | 0.1179651 |
| C | C1 | 782 | 0.1254824 |
|  | C2 | 707 | 0.1298427 |
|  | C3 | 670 | 0.1330496 |
|  | C4 | 532 | 0.1351445 |
|  | C5 | 492 | 0.1386614 |
| D | D1 | 422 | 0.1422396 |
|  | D2 | 465 | 0.1462951 |
|  | D3 | 432 | 0.1506917 |
|  | D4 | 326 | 0.1534770 |
|  | D5 | 251 | 0.1577833 |
| E | E1 | 205 | 0.1606790 |
|  | E2 | 177 | 0.1634503 |
|  | E3 | 136 | 0.1677059 |
|  | E4 | 107 | 0.1703168 |
|  | E5 | 97 | 0.1748722 |
| F | F1 | 60 | 0.1780300 |
|  | F2 | 58 | 0.1803810 |
|  | F3 | 36 | 0.1859361 |
|  | F4 | 38 | 0.1883421 |
|  | F5 | 17 | 0.1903706 |
| G | G1 | 16 | 0.1963000 |
|  | G2 | 13 | 0.2013077 |
|  | G3 | 10 | 0.2064400 |
|  | G4 | 12 | 0.2085833 |
|  | G5 | 8 | 0.2063250 |

10. AVERAGE TIME BETWEEN LAON ACCEPTANCE AND LOAN ISSUANCE DATE.

**data** diff;

set final2;

diff\_ai=sum(issue\_d,-accept\_d);

**run**;

**proc** **means** mean data =diff;

var diff\_ai;

**run**;

OUTPUT:

|  |
| --- |
| **Analysis Variable: diff\_ai** |
| **Mean** |
| 9.3840635 |

**11.** There is no customer whose data is available in loan dataset but not in demographic dataset. There is no customer whose data is available in demographic dataset but not in loan dataset.

**proc** **sort** data=combine;

by id;

**run**;

**proc** **sort** data=demo;

by id;

**run**;

**data** loan\_demo;

merge combine(in=a) demo(in=b);

by id;

if a and not b;

**run**;

There is no customer whose data were available in demographic dataset but not in loan dataset.

**data** demo\_loan;

merge combine(in=a)demo(in=b);

by id;

if b and not a;

**run**;

**12.** There are three important finding from above Questions.

**1st finding:** on the basis of above Q6 average grant rate vary with income level.

**proc** **plot** data= final3;

plot annual\_inc\*avg\_grant\_rate ='\*';

**run**;

The PLOT Procedure

Plot of annual\_inc\*avg\_grant\_rate. Symbol used is '\*'.

6 + \*

|

|

|

| \*

5 +

|

|

a |

n |

n 4 + \*

u |

a |

l |

\_ |

i 3 + \*

n |

c |

|

|

2 + \*

|

|

|

|

1 + \*

-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

. . . . . . . . . . . . . . . . . . . . . . . . . .

0 0 0 0 0 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4 5

0 2 4 6 8 0 2 4 6 8 0 2 4 6 8 0 2 4 6 8 0 2 4 6 8 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

avg\_grant\_rate > Here it’s evident from plot that people of income group 6 having lowest avg\_grant\_rate(Loan requested to loan granted) this mean banks are most prefer to this income group to grant Loan .

* And Income group5 having highest avg\_grant\_rate means Banks grant lowest amount of Loan against Loan requested amount. Reason should must be figure out?

**2ND finding: I**mpact of income verification status on default rate.

**proc** **freq** data=final2;

table is\_inc\_v\*loan\_status/nocol norow;

run;

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table of is\_inc\_v by loan\_status** | | | | | | | | |
| **is\_inc\_v (is\_inc\_v)** | **loan\_status (loan\_status)** | | | | | | | **Total** |
| **Frequency Percent** | **Charged Off** | **Current** | **Default** | **Fully Paid** | **In Grace Period** | **Late (16-30 days)** | **Late (31-120 days)** |
| **1** | 404 3.08 | 666 5.08 | 2 0.02 | 2099 16.02 | 16 0.12 | 4 0.03 | 22 0.17 | 3213 24.52 |
| **0** | 1113 8.49 | 1341 10.24 | 4 0.03 | 7326 55.92 | 26 0.20 | 6 0.05 | 73 0.56 | 9889 75.48 |
| **Total** | **1517 11.58** | **2007 15.32** | **6 0.05** | **9425 71.94** | **42 0.32** | **10 0.08** | **95 0.73** | **13102 100.00** |

According default dataset false status of verification variable(is\_inc\_v) has high percentage(0.0305296901) of default.

There are four loans defaulted when Income verification variable has false status where there are only 2 loan defaulted when verification variable has true status.

So there is huge impact of income verification on default rate of loan.

**3rd finding:** on the basis of Q7 we found that state CA, NY is overrated n diff between loan contribution and population is more than 6.5% and 3.26%and some states are underrated like MI,AZ,TN. Therefore reason should be find out?

**proc** **plot** data= loan\_population\_comp;

plot ratio\_pop\*loan\_percentage =states ;

**run**;

The PLOT Procedure

Plot of ratio\_pop\*loan\_percentage. Symbol is value of States.

ratio\_pop |

|

14 +

|

13 +

|

12 + C

|

11 +

|

10 +

|

9 +

| T

8 +

|

7 +

|

6 + F N

|

5 +

|

4 + I

| O

3 + N M G N

| V

2 +I W M W M M

| KLASM C

1 +IKAOO C

|NWN

0 +VA D

-+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+--

0 1.5 3 4.5 6 7.5 9 10.5 12 13.5 15 16.5 18 19.5

loan\_percentage

15 Observations are hidden.