In [1]:
```
library(dplyr)
library(tidyr)
library(ggplot2)
```

```
Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union
```

In [2]:
```
df <- read.csv("lending-club-loan-data/loan.csv")
head(df)
```

id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installr
1077501	1296599	5000	5000	4975	36 months	10.65	162.87
1077430	1314167	2500	2500	2500	60 months	15.27	59.83
1077175	1313524	2400	2400	2400	36 months	15.96	84.33
1076863	1277178	10000	10000	10000	36 months	13.49	339.31
1075358	1311748	3000	3000	3000	60 months	12.69	67.79
1075269	1311441	5000	5000	5000	36 months	7.90	156.46

In [71]:
```
bcolnames(df)
```

1. 'id'
2. 'member_id'
3. 'loan_amnt'
4. 'funded_amnt'
5. 'funded_amnt_inv'
6. 'term'
7. 'int_rate'
8. 'installment'
9. 'grade'

10. 'sub_grade'
11. 'emp_title'
12. 'emp_length'
13. 'home_ownership'
14. 'annual_inc'
15. 'verification_status'
16. 'issue_d'
17. 'loan_status'
18. 'pymnt_plan'
19. 'url'
20. 'desc'
21. 'purpose'
22. 'title'
23. 'zip_code'
24. 'addr_state'
25. 'dti'
26. 'delinq_2yrs'
27. 'earliest_cr_line'
28. 'inq_last_6mths'
29. 'mths_since_last_delinq'
30. 'mths_since_last_record'
31. 'open_acc'
32. 'pub_rec'
33. 'revol_bal'
34. 'revol_util'
35. 'total_acc'
36. 'initial_list_status'
37. 'out_prncp'
38. 'out_prncp_inv'
39. 'total_pymnt'
40. 'total_pymnt_inv'
41. 'total_rec_prncp'
42. 'total_rec_int'
43. 'total_rec_late_fee'
44. 'recoveries'
45. 'collection_recovery_fee'
46. 'last_pymnt_d'
47. 'last_pymnt_amnt'
48. 'next_pymnt_d'
49. 'last_credit_pull_d'
50. 'collections_12_mths_ex_med'
51. 'mths_since_last_major_derog'
52. 'policy_code'
53. 'application_type'

54. 'annual_inc_joint'
55. 'dti_joint'
56. 'verification_status_joint'
57. 'acc_now_delinq'
58. 'tot_coll_amt'
59. 'tot_cur_bal'
60. 'open_acc_6m'
61. 'open_il_6m'
62. 'open_il_12m'
63. 'open_il_24m'
64. 'mths_since_rcnt_il'
65. 'total_bal_il'
66. 'il_util'
67. 'open_rv_12m'
68. 'open_rv_24m'
69. 'max_bal_bc'
70. 'all_util'
71. 'total_rev_hi_lim'
72. 'inq_fi'
73. 'total_cu_tl'
74. 'inq_last_12m'

In [ ]:

In [3]:
```
summary(is.na(df))
```

```
     id            member_id       loan_amnt       funded_amnt
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:887379    FALSE:887379    FALSE:887379    FALSE:887379

 funded_amnt_inv    term           int_rate       installment
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:887379    FALSE:887379    FALSE:887379    FALSE:887379

   grade          sub_grade       emp_title       emp_length
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:887379    FALSE:887379    FALSE:887379    FALSE:887379

 home_ownership  annual_inc      verification_status  issue_d
 Mode :logical   Mode :logical   Mode :logical       Mode :logical
 FALSE:887379    FALSE:887375    FALSE:887379        FALSE:887379
                 TRUE :4
 loan_status     pymnt_plan         url             desc
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:887379    FALSE:887379    FALSE:887379    FALSE:887379

  purpose          title          zip_code       addr_state
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
```

```
 FALSE:887379    FALSE:887379    FALSE:887379    FALSE:887379

    dti           delinq_2yrs     earliest_cr_line inq_last_6mths
 Mode :logical   Mode :logical   Mode :logical    Mode :logical
 FALSE:887379    FALSE:887350    FALSE:887379     FALSE:887350
                 TRUE :29                         TRUE :29
 mths_since_last_delinq mths_since_last_record  open_acc        pub_rec
 Mode :logical          Mode :logical          Mode :logical   Mode :lo
gical
 FALSE:433067           FALSE:137053           FALSE:887350    FALSE:88
7350
 TRUE :454312           TRUE :750326           TRUE :29        TRUE :29
 revol_bal       revol_util      total_acc       initial_list_status
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:887379    FALSE:886877    FALSE:887350    FALSE:887379
                 TRUE :502       TRUE :29
 out_prncp       out_prncp_inv   total_pymnt     total_pymnt_inv
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:887379    FALSE:887379    FALSE:887379    FALSE:887379

 total_rec_prncp total_rec_int   total_rec_late_fee recoveries
 Mode :logical   Mode :logical   Mode :logical      Mode :logical
 FALSE:887379    FALSE:887379    FALSE:887379       FALSE:887379

 collection_recovery_fee last_pymnt_d    last_pymnt_amnt next_pymnt_d
 Mode :logical           Mode :logical   Mode :logical   Mode :logical
 FALSE:887379            FALSE:887379    FALSE:887379    FALSE:887379

 last_credit_pull_d collections_12_mths_ex_med mths_since_last_major_de
rog
 Mode :logical      Mode :logical              Mode :logical
 FALSE:887379       FALSE:887234               FALSE:221703
                    TRUE :145                  TRUE :665676
 policy_code     application_type annual_inc_joint dti_joint
 Mode :logical   Mode :logical    Mode :logical    Mode :logical
 FALSE:887379    FALSE:887379     FALSE:511        FALSE:509
                                  TRUE :886868     TRUE :886870
 verification_status_joint acc_now_delinq  tot_coll_amt    tot_cur_bal
 Mode :logical             Mode :logical   Mode :logical   Mode :logica
l
 FALSE:887379              FALSE:887350    FALSE:817103    FALSE:817103
                           TRUE :29        TRUE :70276     TRUE :70276
 open_acc_6m     open_il_6m      open_il_12m     open_il_24m
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:21372     FALSE:21372     FALSE:21372     FALSE:21372
 TRUE :866007    TRUE :866007    TRUE :866007    TRUE :866007
 mths_since_rcnt_il total_bal_il     il_util        open_rv_12m
 Mode :logical      Mode :logical   Mode :logical   Mode :logical
 FALSE:20810        FALSE:21372     FALSE:18617     FALSE:21372
 TRUE :866569       TRUE :866007    TRUE :868762    TRUE :866007
 open_rv_24m     max_bal_bc       all_util       total_rev_hi_lim
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
```

```
FALSE:21372     FALSE:21372     FALSE:21372     FALSE:817103
TRUE :866007    TRUE :866007    TRUE :866007    TRUE :70276
  inq_fi         total_cu_tl     inq_last_12m
Mode :logical   Mode :logical   Mode :logical
FALSE:21372     FALSE:21372     FALSE:21372
TRUE :866007    TRUE :866007    TRUE :866007
```

In [4]:
```
too_many_na <- function(x) sum(is.na(x)) < 100000
```

In [5]:
```
df2 <- df %>% select_if(too_many_na)
head(df2)
```

id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installr
1077501	1296599	5000	5000	4975	36 months	10.65	162.87
1077430	1314167	2500	2500	2500	60 months	15.27	59.83
1077175	1313524	2400	2400	2400	36 months	15.96	84.33
1076863	1277178	10000	10000	10000	36 months	13.49	339.31
1075358	1311748	3000	3000	3000	60 months	12.69	67.79
1075269	1311441	5000	5000	5000	36 months	7.90	156.46

In [6]:
```
summary(is.na(df2))
```

```
    id           member_id       loan_amnt       funded_amnt
Mode :logical   Mode :logical   Mode :logical   Mode :logical
FALSE:887379    FALSE:887379    FALSE:887379    FALSE:887379

funded_amnt_inv    term          int_rate       installment
Mode :logical   Mode :logical   Mode :logical   Mode :logical
FALSE:887379    FALSE:887379    FALSE:887379    FALSE:887379

  grade          sub_grade       emp_title       emp_length
Mode :logical   Mode :logical   Mode :logical   Mode :logical
FALSE:887379    FALSE:887379    FALSE:887379    FALSE:887379

home_ownership  annual_inc      verification_status  issue_d
Mode :logical   Mode :logical   Mode :logical       Mode :logical
FALSE:887379    FALSE:887375    FALSE:887379        FALSE:887379
                TRUE :4
loan_status     pymnt_plan         url             desc
Mode :logical   Mode :logical   Mode :logical   Mode :logical
```

```
 FALSE:887379    FALSE:887379    FALSE:887379    FALSE:887379

  purpose          title           zip_code       addr_state
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:887379    FALSE:887379    FALSE:887379    FALSE:887379

    dti          delinq_2yrs     earliest_cr_line inq_last_6mths
 Mode :logical   Mode :logical   Mode :logical    Mode :logical
 FALSE:887379    FALSE:887350    FALSE:887379     FALSE:887350
                 TRUE :29                         TRUE :29
  open_acc        pub_rec        revol_bal       revol_util
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:887350    FALSE:887350    FALSE:887379    FALSE:886877
 TRUE :29        TRUE :29                        TRUE :502
 total_acc       initial_list_status out_prncp       out_prncp_inv
 Mode :logical   Mode :logical       Mode :logical   Mode :logical
 FALSE:887350    FALSE:887379        FALSE:887379    FALSE:887379
 TRUE :29
 total_pymnt     total_pymnt_inv total_rec_prncp total_rec_int
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:887379    FALSE:887379    FALSE:887379    FALSE:887379

 total_rec_late_fee recoveries      collection_recovery_fee last_pymnt_
d
 Mode :logical      Mode :logical   Mode :logical           Mode :logic
al
 FALSE:887379       FALSE:887379    FALSE:887379            FALSE:88737
9

 last_pymnt_amnt next_pymnt_d    last_credit_pull_d collections_12_mths
_ex_med
 Mode :logical   Mode :logical   Mode :logical      Mode :logical
 FALSE:887379    FALSE:887379    FALSE:887379       FALSE:887234
                                                    TRUE :145
 policy_code     application_type verification_status_joint acc_now_del
inq
 Mode :logical   Mode :logical    Mode :logical             Mode :logic
al
 FALSE:887379    FALSE:887379     FALSE:887379              FALSE:88735
0
                                                           TRUE :29
 tot_coll_amt    tot_cur_bal     total_rev_hi_lim
 Mode :logical   Mode :logical   Mode :logical
 FALSE:817103    FALSE:817103    FALSE:817103
 TRUE :70276     TRUE :70276     TRUE :70276
```

In [7]:
```
df2 <- df2 %>% drop_na()
summary(is.na(df2))
```

```
     id           member_id       loan_amnt       funded_amnt
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:816722    FALSE:816722    FALSE:816722    FALSE:816722
 funded_amnt_inv    term          int_rate       installment
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:816722    FALSE:816722    FALSE:816722    FALSE:816722
   grade          sub_grade       emp_title       emp_length
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:816722    FALSE:816722    FALSE:816722    FALSE:816722
 home_ownership  annual_inc      verification_status  issue_d
 Mode :logical   Mode :logical   Mode :logical       Mode :logical
 FALSE:816722    FALSE:816722    FALSE:816722        FALSE:816722
 loan_status     pymnt_plan         url             desc
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:816722    FALSE:816722    FALSE:816722    FALSE:816722
  purpose          title          zip_code       addr_state
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:816722    FALSE:816722    FALSE:816722    FALSE:816722
    dti          delinq_2yrs     earliest_cr_line inq_last_6mths
 Mode :logical   Mode :logical   Mode :logical    Mode :logical
 FALSE:816722    FALSE:816722    FALSE:816722     FALSE:816722
  open_acc        pub_rec        revol_bal       revol_util
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:816722    FALSE:816722    FALSE:816722    FALSE:816722
 total_acc       initial_list_status out_prncp       out_prncp_inv
 Mode :logical   Mode :logical       Mode :logical   Mode :logical
 FALSE:816722    FALSE:816722        FALSE:816722    FALSE:816722
 total_pymnt     total_pymnt_inv total_rec_prncp total_rec_int
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:816722    FALSE:816722    FALSE:816722    FALSE:816722
 total_rec_late_fee recoveries      collection_recovery_fee last_pymnt_
d
 Mode :logical      Mode :logical   Mode :logical           Mode :logic
al
 FALSE:816722       FALSE:816722    FALSE:816722            FALSE:81672
2
 last_pymnt_amnt next_pymnt_d    last_credit_pull_d collections_12_mths
_ex_med
 Mode :logical   Mode :logical   Mode :logical      Mode :logical
 FALSE:816722    FALSE:816722    FALSE:816722       FALSE:816722
 policy_code     application_type verification_status_joint acc_now_del
inq
 Mode :logical   Mode :logical    Mode :logical             Mode :logic
al
 FALSE:816722    FALSE:816722     FALSE:816722              FALSE:81672
2
 tot_coll_amt    tot_cur_bal     total_rev_hi_lim
 Mode :logical   Mode :logical   Mode :logical
 FALSE:816722    FALSE:816722    FALSE:816722
```

In [8]:
```
df2 <- df2 %>% filter(application_type == 'INDIVIDUAL')
nrow(df2)/nrow(df)
```

0.919799769884119

In [9]:
```
colnames(df2)
```

1. 'id'
2. 'member_id'
3. 'loan_amnt'
4. 'funded_amnt'
5. 'funded_amnt_inv'
6. 'term'
7. 'int_rate'
8. 'installment'
9. 'grade'
10. 'sub_grade'
11. 'emp_title'
12. 'emp_length'
13. 'home_ownership'
14. 'annual_inc'
15. 'verification_status'
16. 'issue_d'
17. 'loan_status'
18. 'pymnt_plan'
19. 'url'
20. 'desc'
21. 'purpose'
22. 'title'
23. 'zip_code'
24. 'addr_state'
25. 'dti'
26. 'delinq_2yrs'
27. 'earliest_cr_line'
28. 'inq_last_6mths'
29. 'open_acc'
30. 'pub_rec'
31. 'revol_bal'
32. 'revol_util'
33. 'total_acc'
34. 'initial_list_status'
35. 'out_prncp'
36. 'out_prncp_inv'
37. 'total_pymnt'
38. 'total_pymnt_inv'

39. 'total_rec_prncp'
40. 'total_rec_int'
41. 'total_rec_late_fee'
42. 'recoveries'
43. 'collection_recovery_fee'
44. 'last_pymnt_d'
45. 'last_pymnt_amnt'
46. 'next_pymnt_d'
47. 'last_credit_pull_d'
48. 'collections_12_mths_ex_med'
49. 'policy_code'
50. 'application_type'
51. 'verification_status_joint'
52. 'acc_now_delinq'
53. 'tot_coll_amt'
54. 'tot_cur_bal'
55. 'total_rev_hi_lim'

In [10]:
```
unique(df2$loan_status)
```

1. Current
2. Fully Paid
3. Late (31-120 days)
4. Late (16-30 days)
5. Charged Off
6. In Grace Period
7. Default
8. Issued

In [20]:
```
df2$loan_status <- as.character(df2$loan_status)
```

In [61]:
```
head(df2$loan_status, 20)
```

1. 'Current'
2. 'Current'
3. 'Current'
4. 'Fully Paid'
5. 'Current'
6. 'Current'
7. 'Current'
8. 'Current'
9. 'Current'
10. 'Fully Paid'
11. 'Current'
12. 'Fully Paid'
13. 'Current'
14. 'Current'
15. 'Fully Paid'
16. 'Late (31-120 days)'
17. 'Late (16-30 days)'
18. 'Fully Paid'
19. 'Current'
20. 'Fully Paid'

In [23]:
```
print(unique(df2$loan_status))
```

```
[1] "Current"            "Fully Paid"         "Late (31-120 days)"
[4] "Late (16-30 days)"  "Charged Off"        "In Grace Period"
[7] "Default"            "Issued"
```

In [59]:
```
df2$loan_rate <- ifelse(df2$loan_status == "Current" | df2$loan_status
== "Issued", 2,
                        ifelse(df2$loan_status == "Fully Paid", 1, 0))
```

In [60]:
```
head(df2$loan_rate, 20)
```

1. 2
2. 2
3. 2
4. 1
5. 2
6. 2
7. 2
8. 2
9. 2
10. 1
11. 2
12. 1
13. 2
14. 2
15. 1
16. 0
17. 0
18. 1
19. 2
20. 1

In [63]:
```
write.csv(df2, "cleaned_loan_data.csv")
```

In [65]:
```
sum(df2$loan_rate == 0)
```

56376

In [66]:
```
sum(df2$loan_rate == 1)
```

153936

In [67]:
```
sum(df2$loan_rate == 2)
```

605899

In [70]:
```
nrow(df2) == sum(df2$loan_rate == 0) + sum(df2$loan_rate == 1) + sum(df
2$loan_rate == 2)
```

TRUE

In [ ]:
```

```