Midterm_Project Loan

Data Introduction

The dataset contains complete loan data for all loans issued through 2012-2013 stated, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The data can be download:https://resources.lendingclub.com/LoanStats3b.csv.zip

Project Introduction

1.Data manipulation
2.Exploratory Data Analysis
Explore the relationship between loan characteristics,personal income and credit. 3.Modeling
Try to apply regresssion to predict loan status.

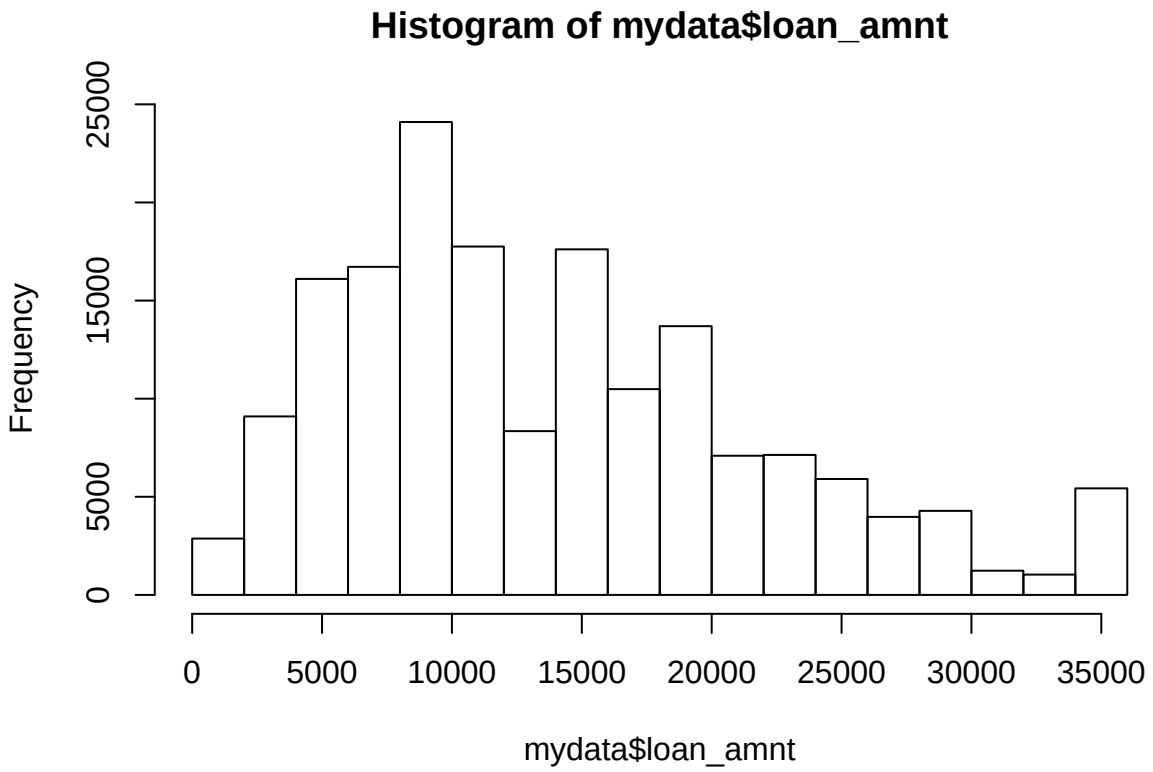
Data Manipulation

```
datap <- select(data,
                id,loan_amnt,loan_status,funded_amnt_inv,term,int_rate,installment,
         grade, annual_inc,verification_status,
   dti,home_ownership,application_type,issue_d
                       )
#For numeric analysis purpose, we need to convert some of the chr object to numeric,
#e.g. interest rate displaying as a charactor '11.5%' need to be converted to 0.115 as a numeric value
mydata<-datap
mydata$application_type<-as.factor(mydata$application_type)
mydata$term <- as.numeric(substr(data$term, 1,3))
mydata$int_rate <- as.numeric(gsub("%", "", data$int_rate)) / 100
#Because of the project capability,I choose the individual application for further analysis.
mydata<-filter(mydata,application_type=="Individual")
#It turns out that all the application type is individual.
```

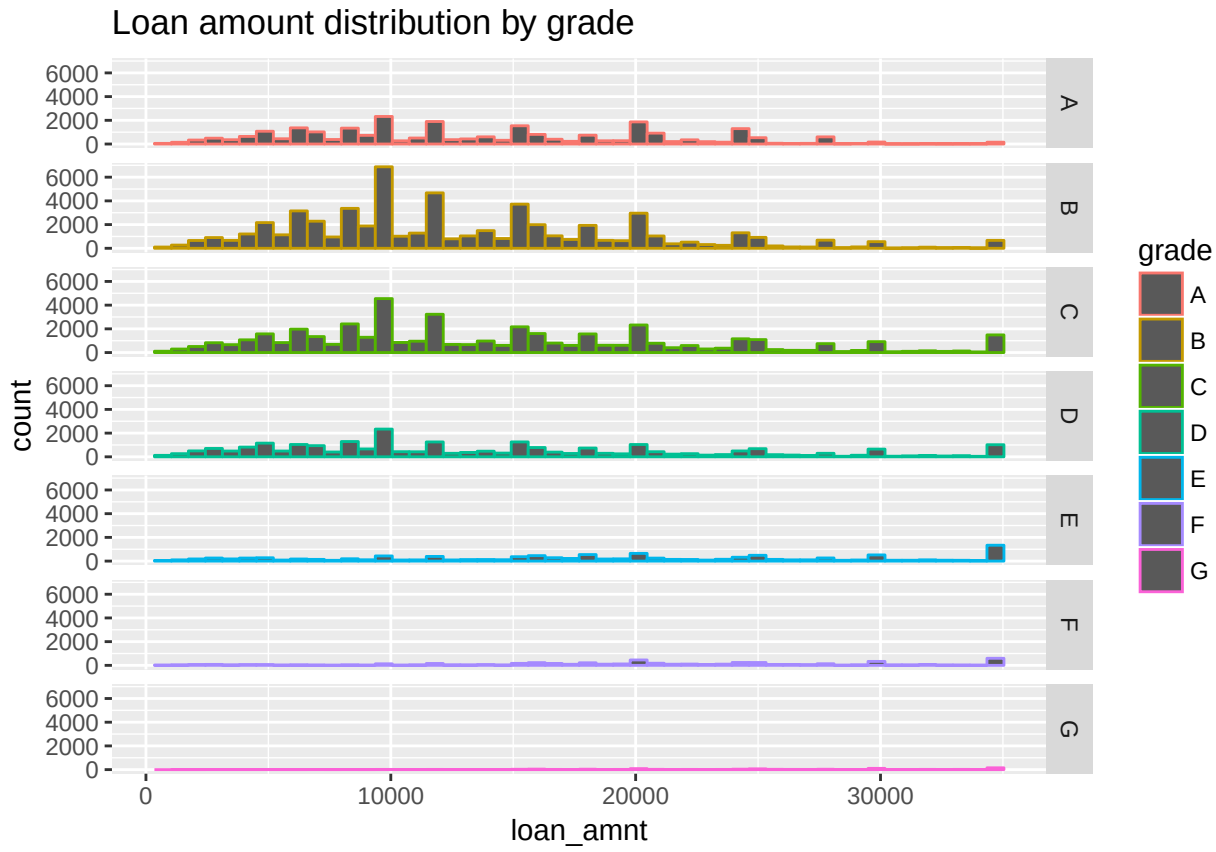
Exploratory Data Analysis

Loan amount distribution

```
hist(mydata$loan_amnt)
```




```
ggplot(mydata, aes(loan_amnt, col = grade)) + geom_histogram(bins = 50) +
  facet_grid(grade ~ .)+labs(title ='Loan amount distribution by grade')
```

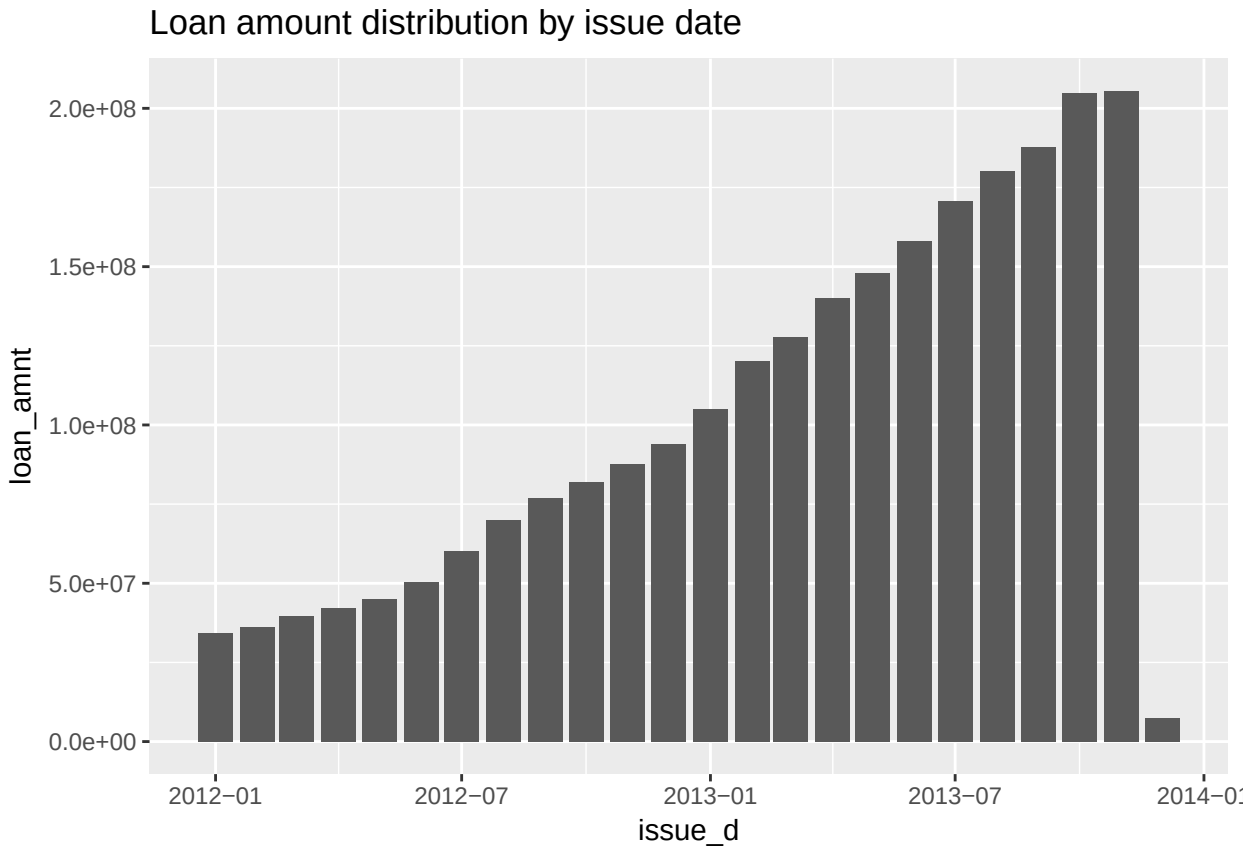


According to the plot,clients with higher grades (A, B, C and D) tend to have received more loans compared to those with lower grades (E, F and G).And for A,B,C and D grades,the most possible loan amount is 10000.

```
mydata$issue_d <- dmy(paste0("01-",mydata$issue_d))
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone 'zone/tz/2017c.1.0/
## zoneinfo/America/New_York'
```

```
loan_amnt_by_month <- aggregate(loan_amnt ~ issue_d, data = mydata, sum)
ggplot(loan_amnt_by_month, aes(issue_d, loan_amnt)) + geom_bar(stat = "identity")+labs(title ='Loan amou
```



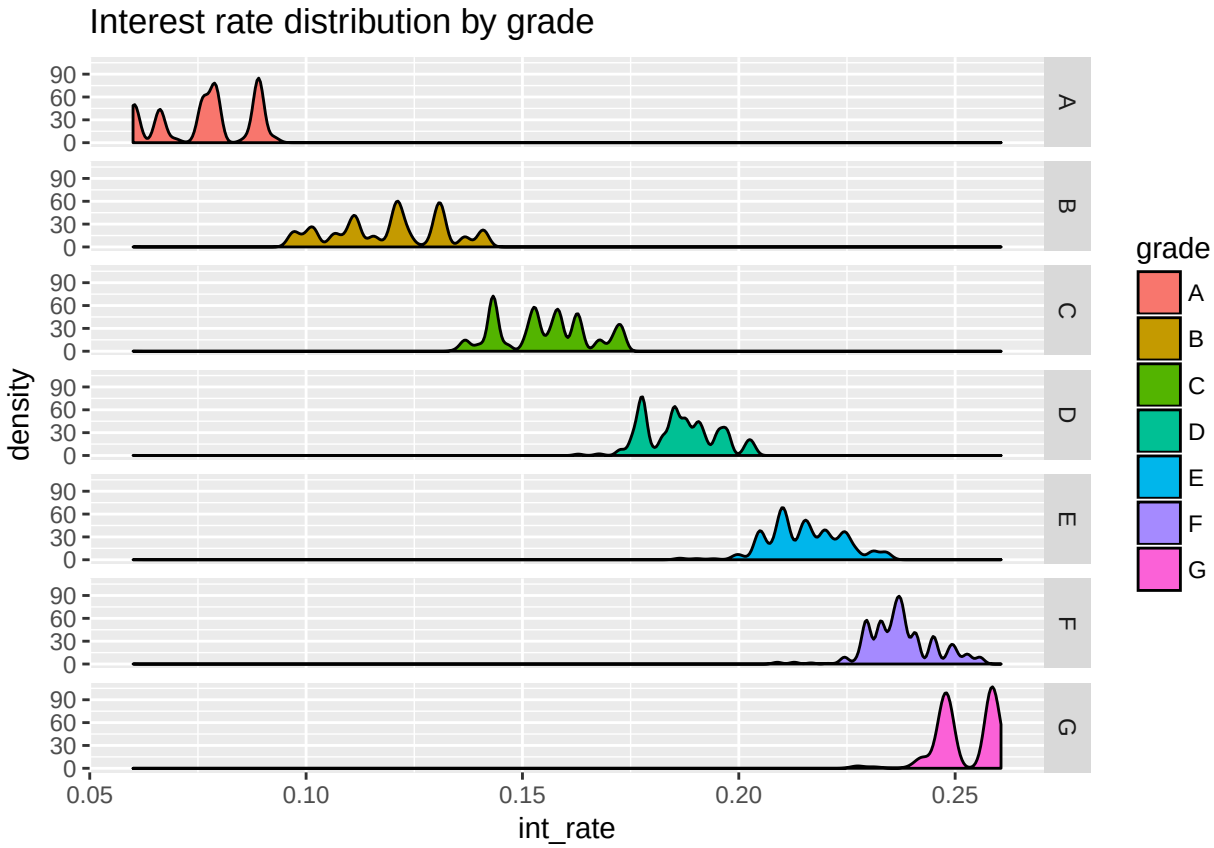

```
colSums(is.na(mydata))
```

```
##                  id           loan_amnt         loan_status
##                   0                   0                   0
##     funded_amnt_inv                term            int_rate
##                   0                   0                   0
##         installment               grade          annual_inc
##                   0                   0                   0
## verification_status                 dti      home_ownership
##                   0                   0                   0
##    application_type             issue_d
##                   0                   0
```

According to the plot,the loan amount increases steadily as time goes,except for the period near 2014,which is probably the consequence of lacking loan amount valid until or after the last period of 2013-2014.

Interest rate

```
ggplot(mydata, aes(int_rate)) + geom_density(aes(fill = grade)) + facet_grid(grade ~ .)+
  labs(title ='Interest rate distribution by grade')
```



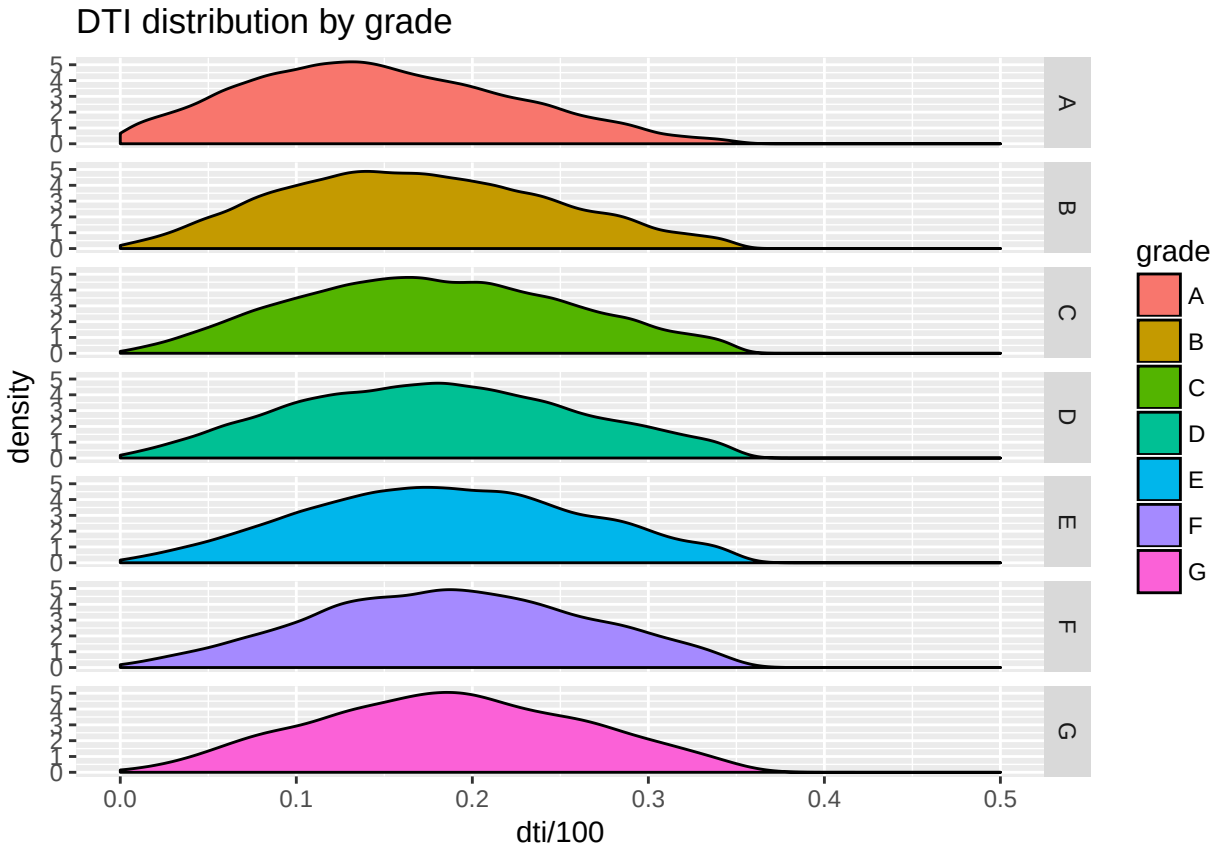
Apparently, interest rates increase as the risk goes up,and the grades are assigned based on risk, so the interest rates changes as the grades changes.

DTI ratio

```
summary(mydata$dti)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   11.32   16.76   17.04   22.55   34.99
```

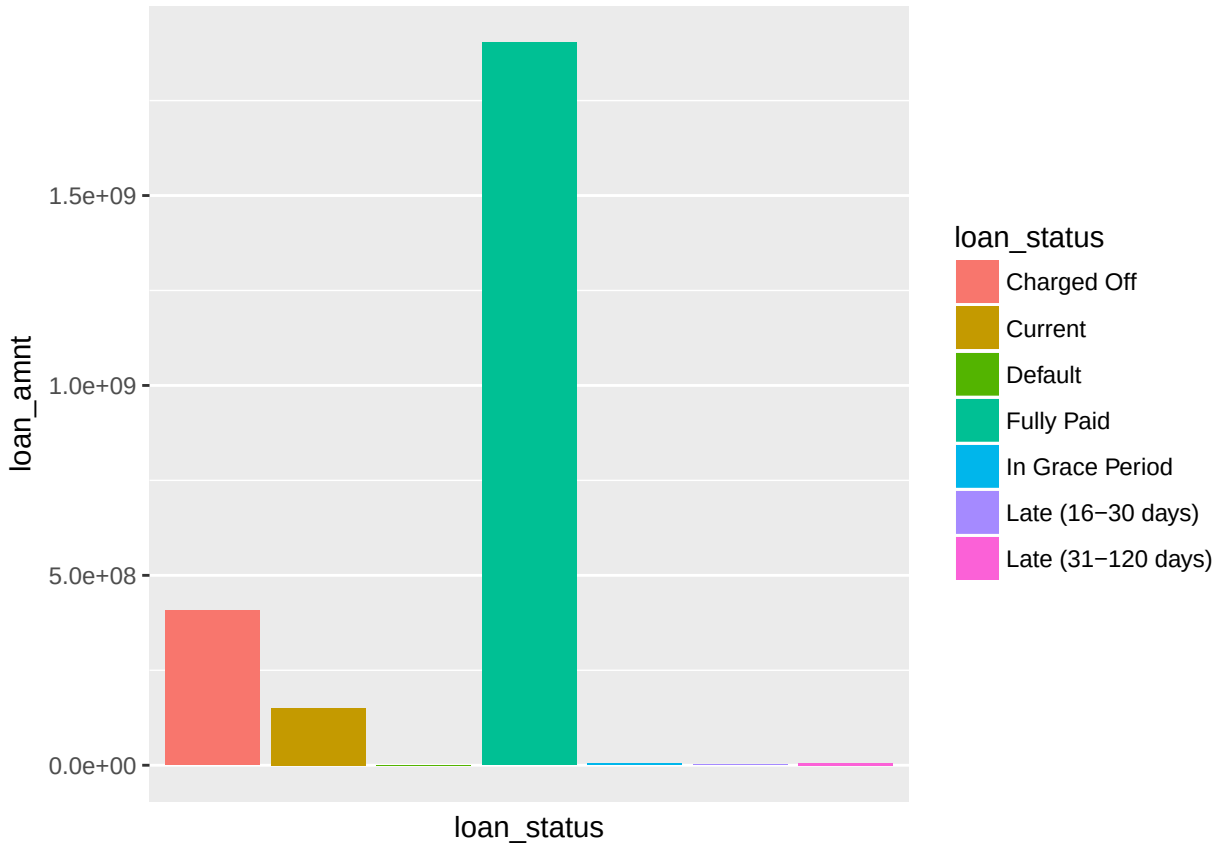
```
ggplot( mydata, aes(dti/100)) + geom_density(aes(fill = grade))+
  facet_grid(grade ~ .)+ xlim(0,0.5) +
  labs(title ='DTI distribution by grade')
```



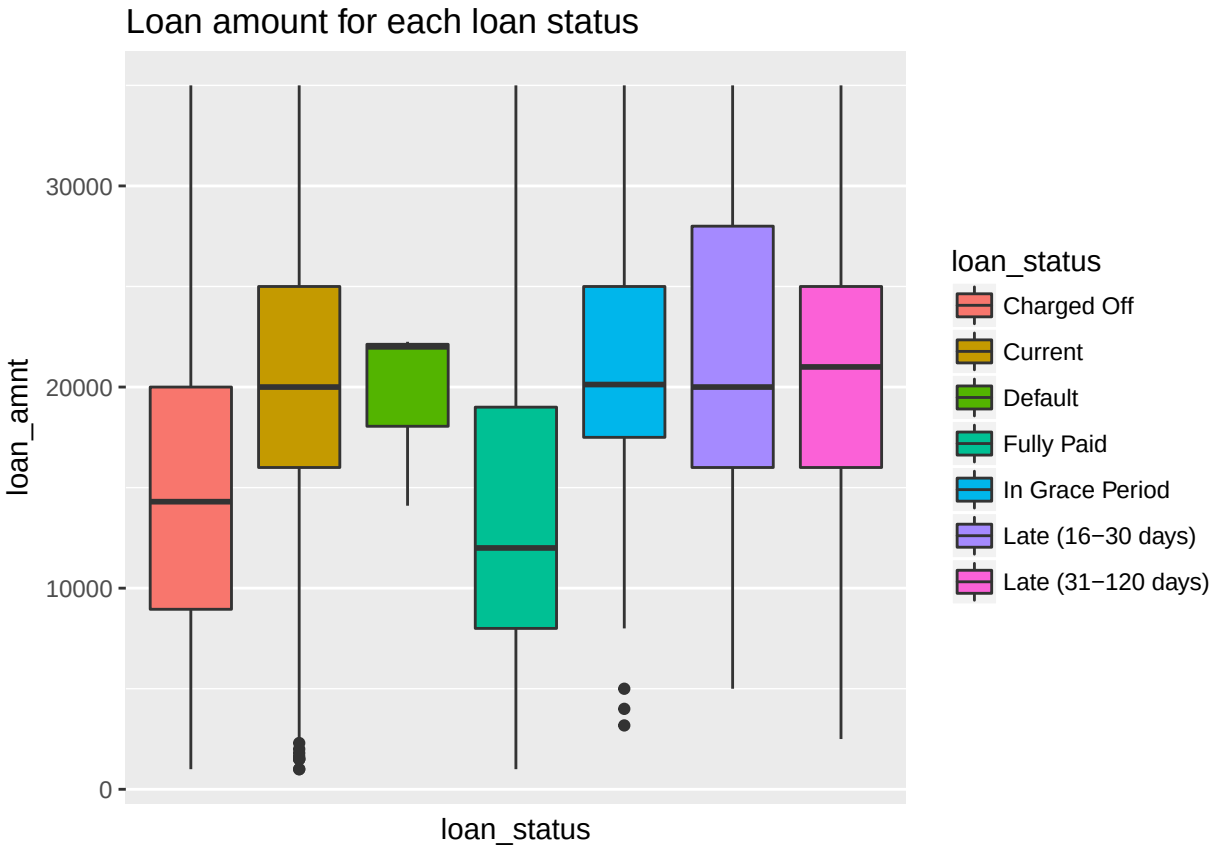
DTI:A ratio calculated using the borrower???s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower???s self-reported monthly income. According to the plot,the DTI ratio tend to increase as the grades go up.For high-grade loan, the majority of the borrowers will not commit more than 20% of the income on debt while low-grade borrowers do tend to loan more.

Total loan amount for each loan status

```
loan_amnt_by_status <- aggregate(loan_amnt ~ loan_status, data = mydata, sum)
ggplot(loan_amnt_by_status, aes(loan_status, loan_amnt, fill = loan_status)) + geom_bar(stat = "identity
```



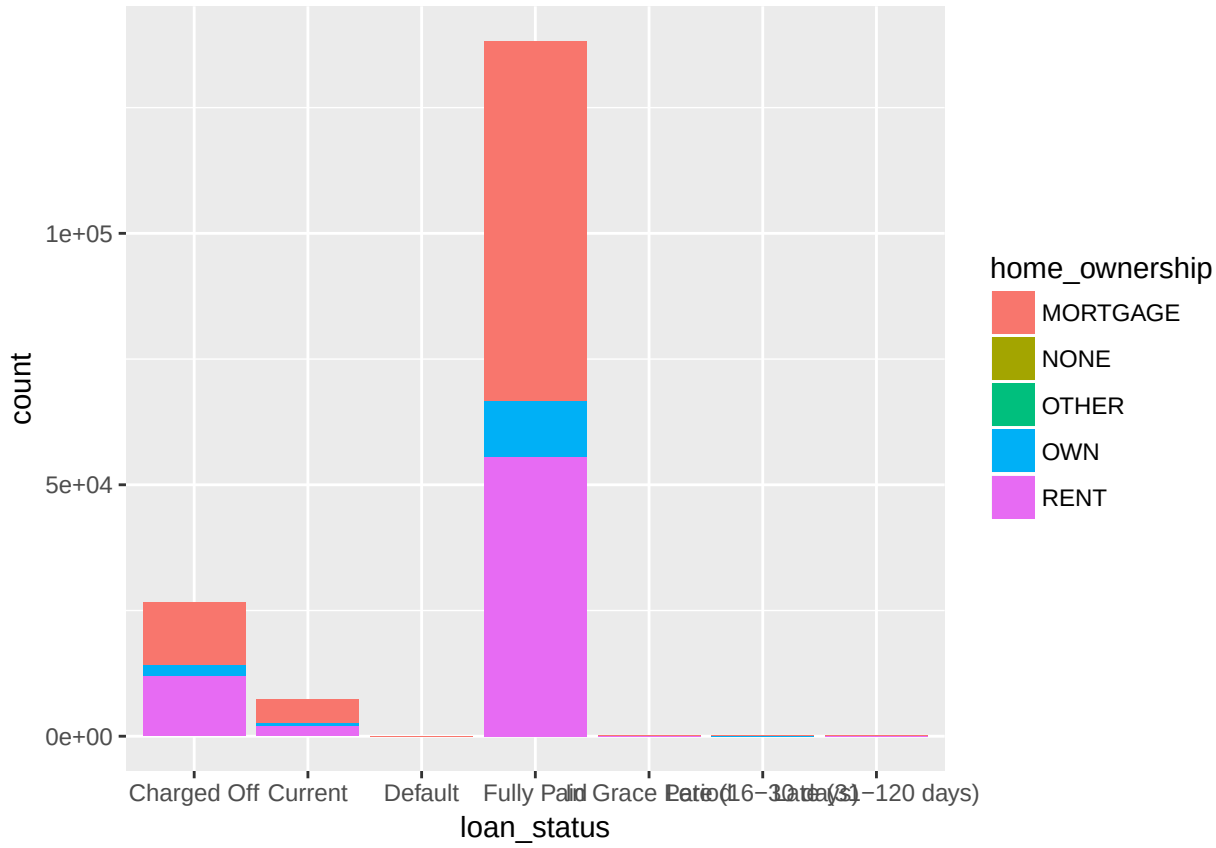

```
#Distribution of the loan amount for each status
ggplot(mydata, aes(loan_status, loan_amnt, fill = loan_status)) + geom_boxplot() + scale_x_discrete(brea
  labs(title ='Loan amount for each loan status')
```



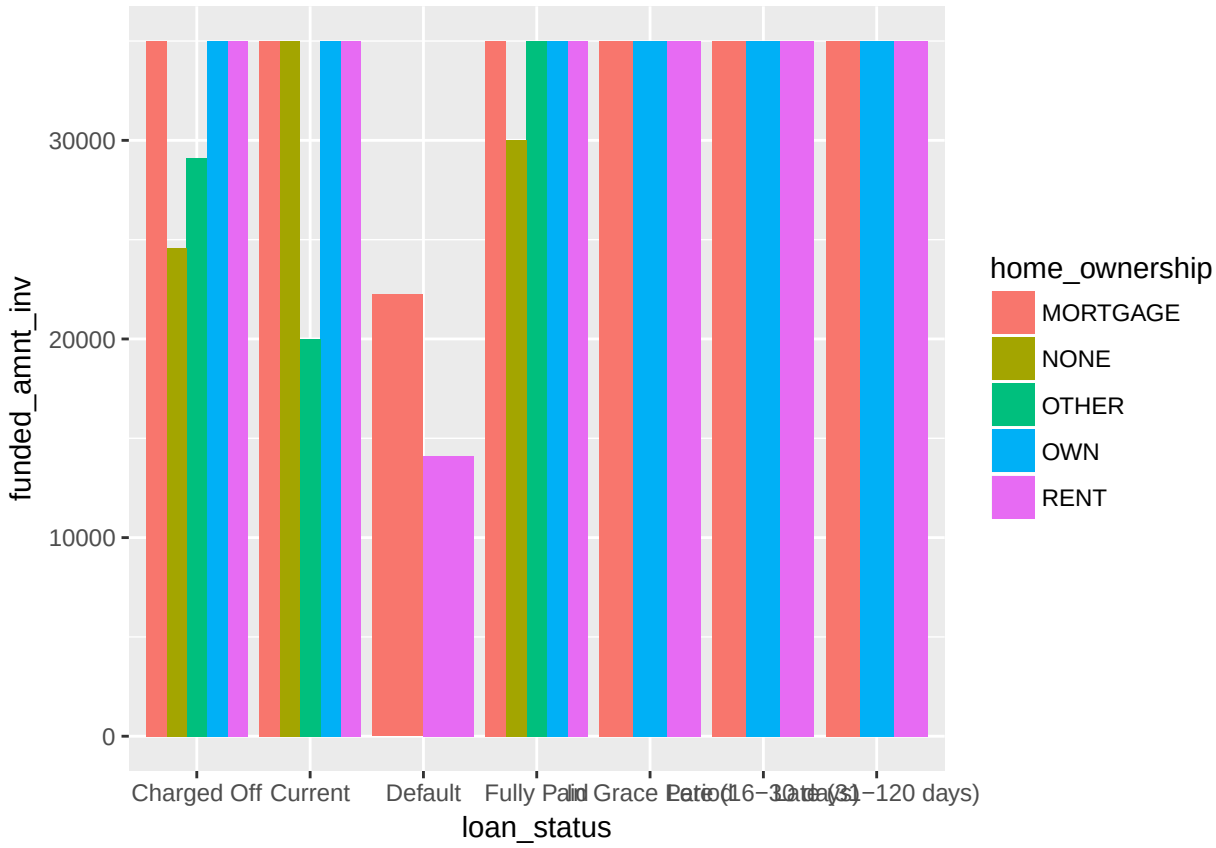
According to the plot,apparently most of loans are fully paid,and there exist differnces of loan amount between loan status.

Home ownership,committed amount and loan status

```
ggplot(data = mydata) +
  geom_bar(mapping = aes(x = loan_status, fill = home_ownership))
```



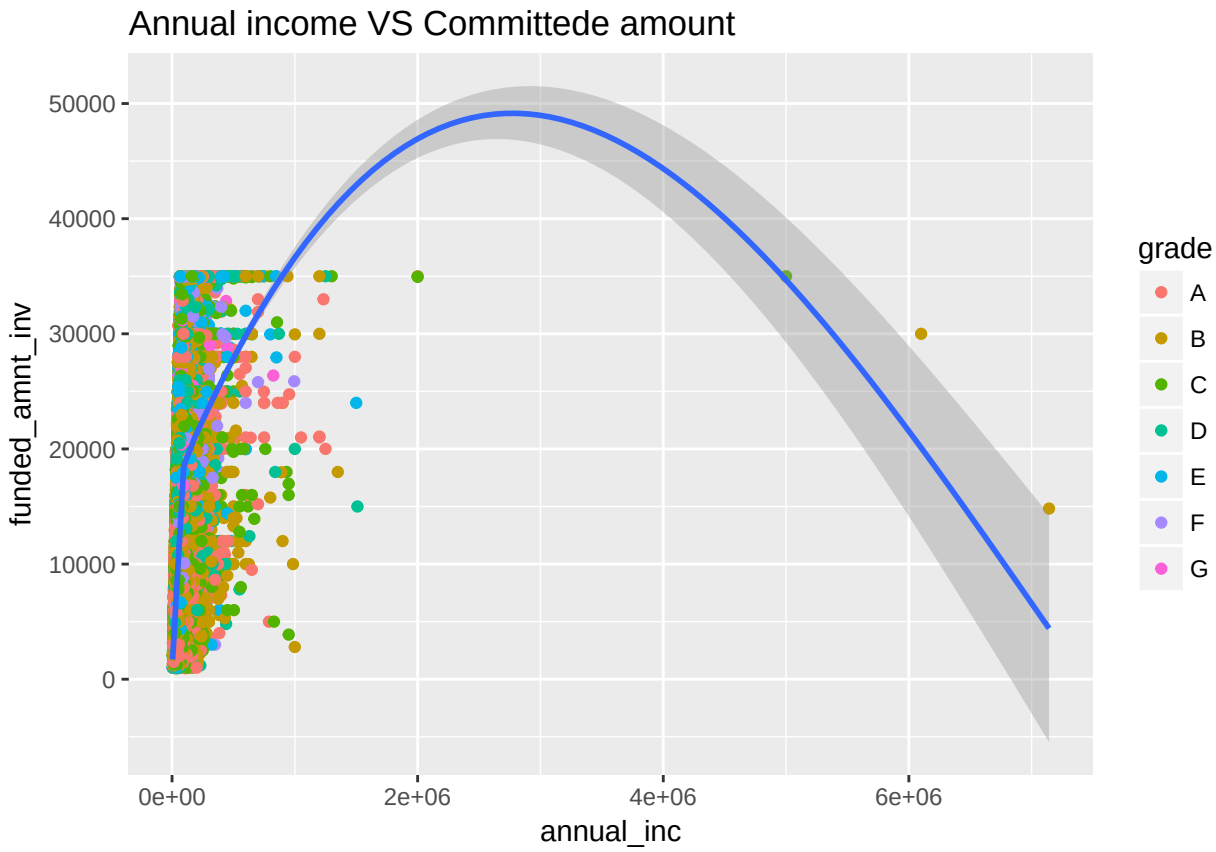

```
ggplot(data = mydata) +
  geom_bar(stat = "identity", mapping=aes(x = loan_status,y =funded_amnt_inv, fill = home_ownership),pos
```



According to the plot,the relationships between loan status,home ownership and total amount committed are complicated.Clients don't own home or other tend to be committed less loan amount by investors.

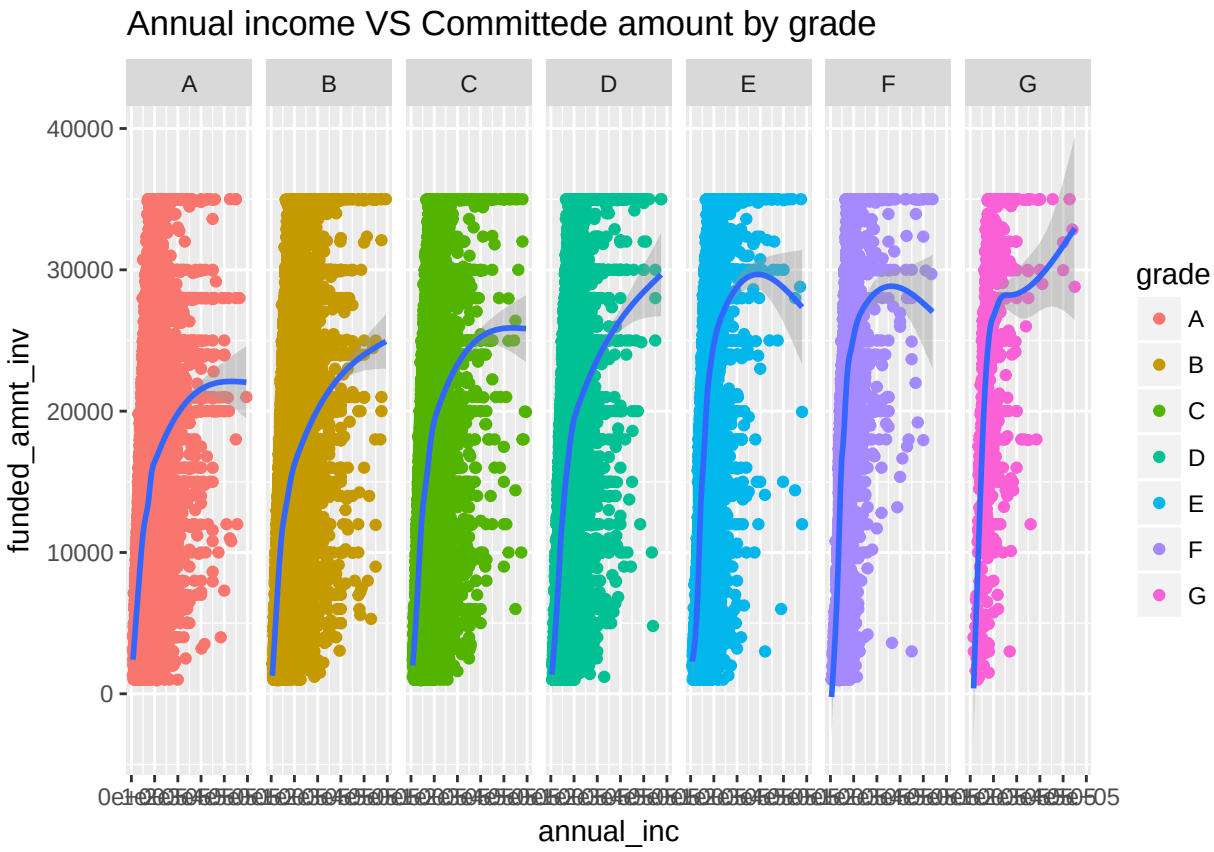
```
ggplot(mydata, aes(annual_inc, funded_amnt_inv)) +
  geom_point(aes(colour = grade))  +
  geom_smooth()+
  labs(title ='Annual income VS Committede amount ')
```

```
## `geom_smooth()` using method = 'gam'
```




```
#when delete the outliers
mydata1 <- filter(mydata, annual_inc < 500000)
#Replot and group by grades
ggplot(mydata1, aes(annual_inc, funded_amnt_inv)) +
  geom_point(aes(colour = grade))  +
  geom_smooth()+ facet_grid(. ~ grade)+
  labs(title ='Annual income VS Committede amount by grade')
```

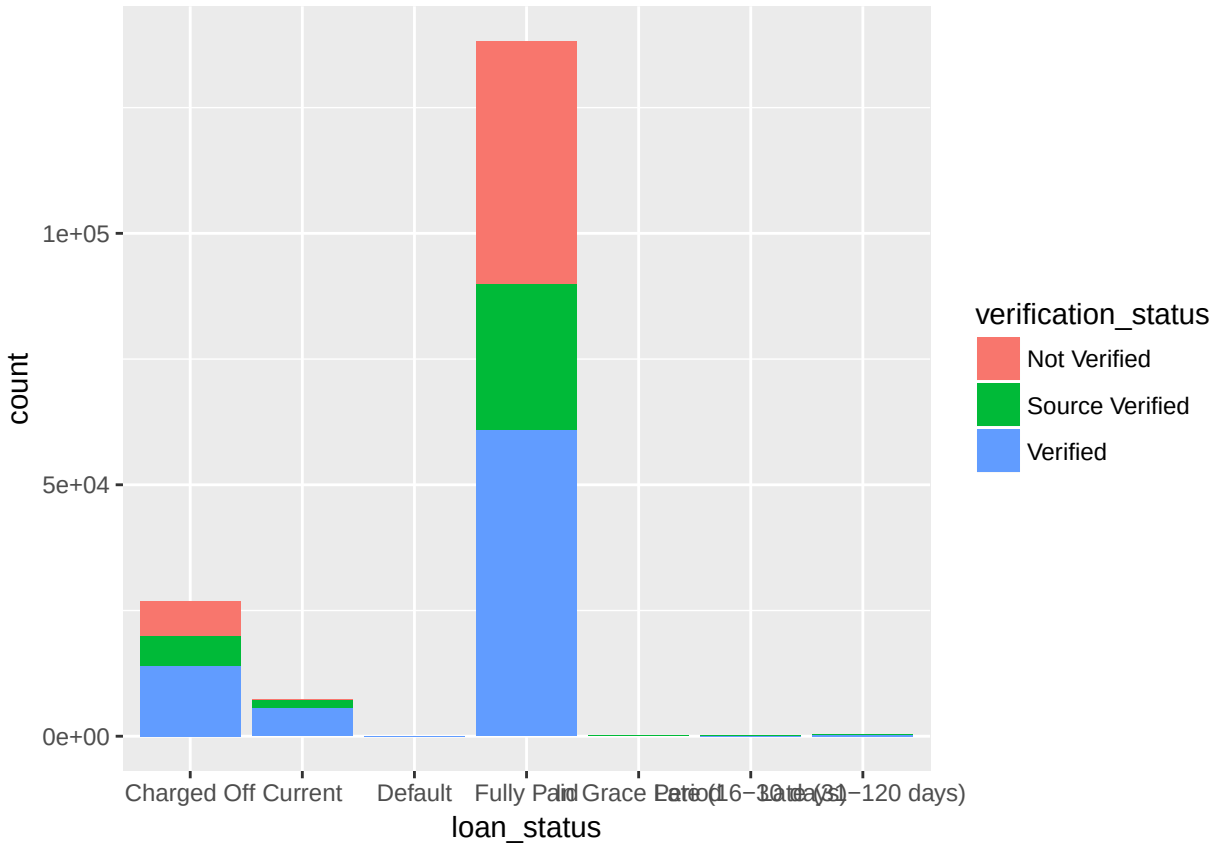
```
## `geom_smooth()` using method = 'gam'
```



According to the plot,there exists definite relationship between income and committed amount,which means cliens with higher income tend to be able to borrow more money.

Verified income

```
ggplot(data = mydata) +
  geom_bar(mapping = aes(x = loan_status, fill = verification_status))
```



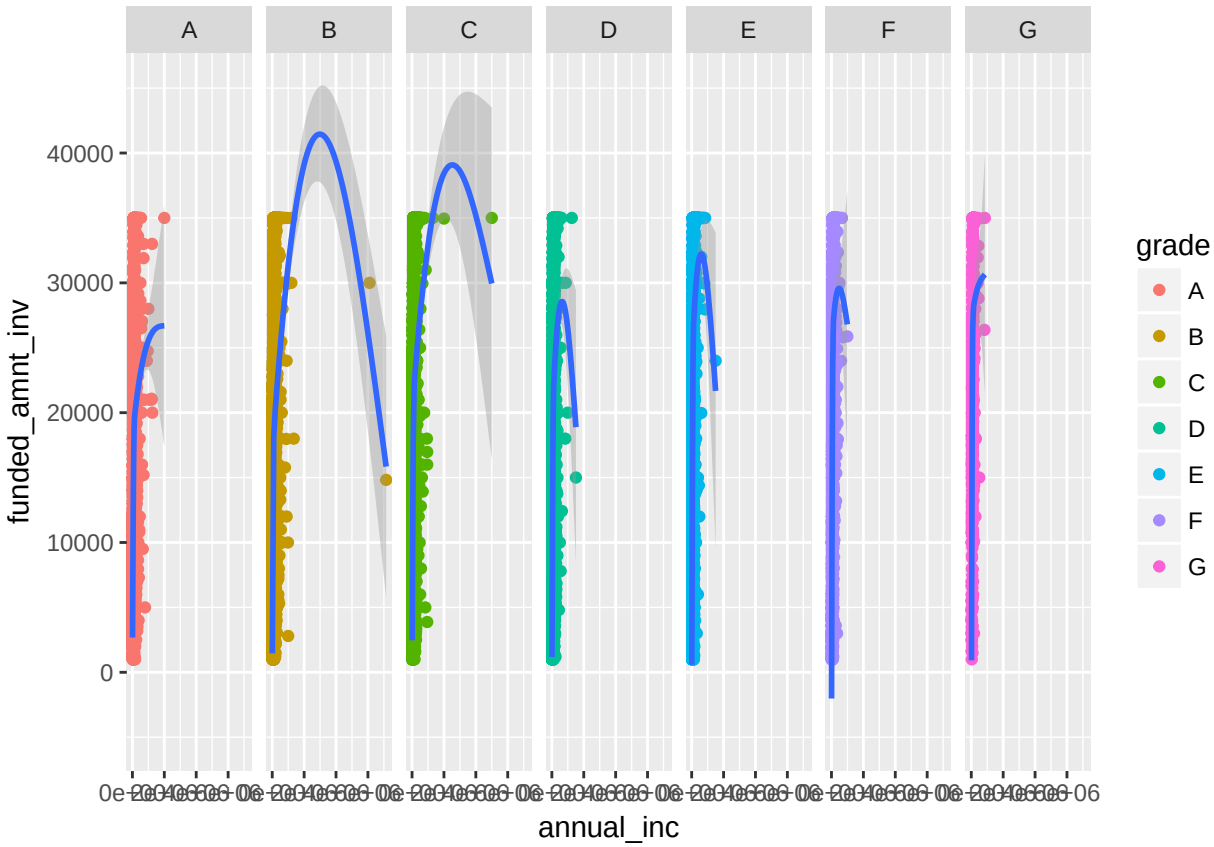
If delete the incomes that are not vertified

```
library(dplyr)
data2= filter(mydata,verification_status!='Not Verified')
nrow(data2)
```

```
## [1] 117882
```

```
#Replot and group by grades
ggplot(data2, aes(annual_inc, funded_amnt_inv)) +
  geom_point(aes(colour = grade))  +
  geom_smooth()+facet_grid(. ~ grade)
```

```
## `geom_smooth()` using method = 'gam'
```



According to the plot there evidently exist relationship between annual income and committed amount for the majority of clients with relatively lower annual income.

Modeling

According to the data and my interest,I'd like to predict the loan status by regression. So I define loan status as paid or not,and labelled as 1 or 0.

```
mydata$loan_paid <- factor(ifelse(mydata$loan_status=="Fully Paid",1,0))
mydata1$loan_paid <- factor(ifelse(mydata1$loan_status=="Fully Paid",1,0))
```

Logistic Model

```
fit1 <- glm(loan_paid~loan_amnt+funded_amnt_inv+term+int_rate+installment+grade+annual_inc+dti+home_own
```

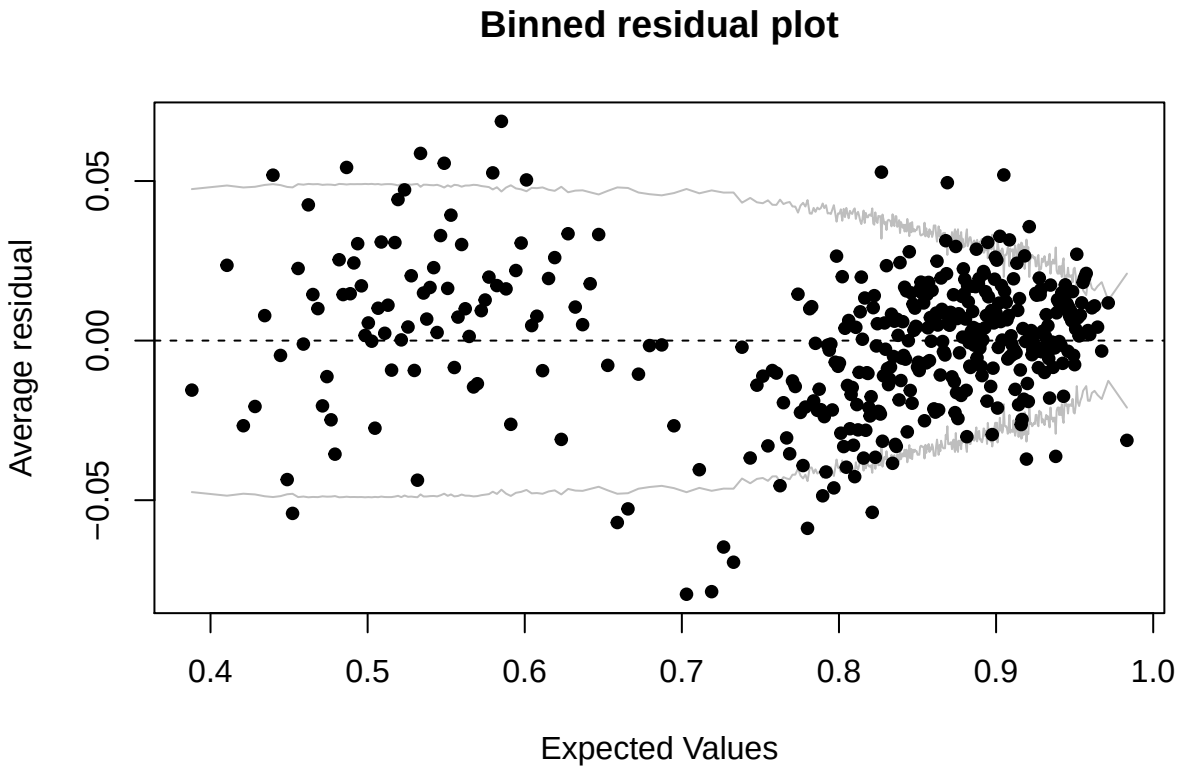
```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(fit1)
```

```
##
## Call:
## glm(formula = loan_paid ~ loan_amnt + funded_amnt_inv + term +
##     int_rate + installment + grade + annual_inc + dti + home_ownership,
##     family = binomial, data = mydata)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -3.4360   0.3281   0.4873   0.6245   1.4597
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          5.609e+00  7.540e-02  74.387  < 2e-16 ***
## loan_amnt           -2.344e-05  2.549e-05  -0.920 0.357759
## funded_amnt_inv      1.798e-05  2.579e-05   0.697 0.485675
## term                -5.640e-02  1.611e-03 -35.003  < 2e-16 ***
## int_rate            -1.030e+01  6.412e-01 -16.070  < 2e-16 ***
## installment         -2.870e-05  1.888e-04  -0.152 0.879182
## gradeB              -1.345e-01  3.880e-02  -3.467 0.000527 ***
## gradeC              -2.038e-01  5.608e-02  -3.634 0.000279 ***
## gradeD              -5.274e-02  7.407e-02  -0.712 0.476436
## gradeE               1.970e-01  9.185e-02   2.145 0.031955 *
## gradeF               3.549e-01  1.067e-01   3.326 0.000881 ***
## gradeG               4.207e-01  1.286e-01   3.271 0.001071 **
## annual_inc           4.538e-06  2.152e-07  21.091  < 2e-16 ***
## dti                 -1.628e-02  8.924e-04 -18.248  < 2e-16 ***
## home_ownershipNONE  -1.272e-01  3.987e-01  -0.319 0.749628
## home_ownershipOTHER -1.683e-01  3.766e-01  -0.447 0.654905
## home_ownershipOWN   -7.145e-02  2.460e-02  -2.905 0.003673 **
## home_ownershipRENT  -1.305e-01  1.440e-02  -9.062  < 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: 173169  on 172874  degrees of freedom
## Residual deviance: 151647  on 172857  degrees of freedom
## AIC: 151683
##
## Number of Fisher Scoring iterations: 5
```

```
binnedplot(fitted(fit1),residuals(fit1,type="response"))
```



According to the binned residual, the model need improvement. According to the EDA,there are interactions can be added to the model.

```
fit2<-glm(loan_paid~loan_amnt+funded_amnt_inv+term+int_rate+installment+grade+annual_inc+dti+home_owners
```

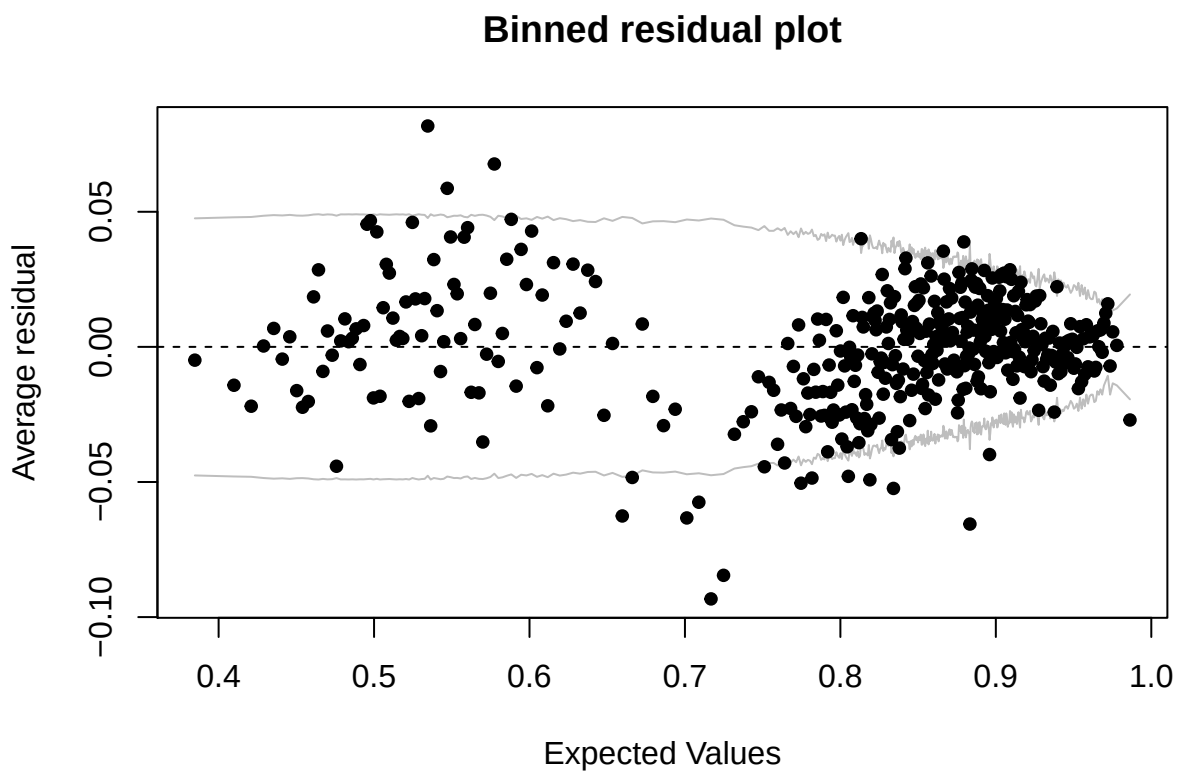
```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(fit2)
```

```
##
## Call:
## glm(formula = loan_paid ~ loan_amnt + funded_amnt_inv + term +
##     int_rate + installment + grade + annual_inc + dti + home_ownership +
##     grade * int_rate, family = binomial, data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4339   0.3058   0.4890   0.6260   1.4858
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          7.367e+00  2.240e-01  32.888  < 2e-16 ***
## loan_amnt           -2.245e-05  2.552e-05  -0.880  0.37913
## funded_amnt_inv      1.704e-05  2.580e-05   0.660  0.50893
## term                -5.629e-02  1.614e-03 -34.869  < 2e-16 ***
## int_rate            -3.253e+01  2.703e+00 -12.035  < 2e-16 ***
## installment         -3.043e-05  1.891e-04  -0.161  0.87215
## gradeB              -2.043e+00  2.500e-01  -8.173 3.01e-16 ***
## gradeC              -1.949e+00  2.789e-01  -6.988 2.79e-12 ***
## gradeD              -3.323e+00  3.903e-01  -8.514  < 2e-16 ***
## gradeE              -1.280e+00  5.494e-01  -2.329  0.01985 *
```

```
## gradeF              -6.263e-01  8.812e-01  -0.711  0.47724
## gradeG              -1.999e+00  2.357e+00  -0.848  0.39625
## annual_inc           4.533e-06  2.152e-07  21.065  < 2e-16 ***
## dti                 -1.615e-02  8.925e-04 -18.098  < 2e-16 ***
## home_ownershipNONE  -1.301e-01  3.981e-01  -0.327  0.74392
## home_ownershipOTHER -1.703e-01  3.762e-01  -0.453  0.65072
## home_ownershipOWN   -7.134e-02  2.460e-02  -2.900  0.00373 **
## home_ownershipRENT  -1.297e-01  1.440e-02  -9.008  < 2e-16 ***
## int_rate:gradeB      2.342e+01  2.882e+00   8.126 4.42e-16 ***
## int_rate:gradeC      2.209e+01  2.918e+00   7.573 3.64e-14 ***
## int_rate:gradeD      3.029e+01  3.207e+00   9.445  < 2e-16 ***
## int_rate:gradeE      2.088e+01  3.571e+00   5.847 5.01e-09 ***
## int_rate:gradeF      1.892e+01  4.498e+00   4.206 2.60e-05 ***
## int_rate:gradeG      2.482e+01  9.684e+00   2.562  0.01039 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 173169  on 172874  degrees of freedom
## Residual deviance: 151552  on 172851  degrees of freedom
## AIC: 151600
##
## Number of Fisher Scoring iterations: 5
```

```
binnedplot(fitted(fit2),residuals(fit2,type="response"))
```



Improved.

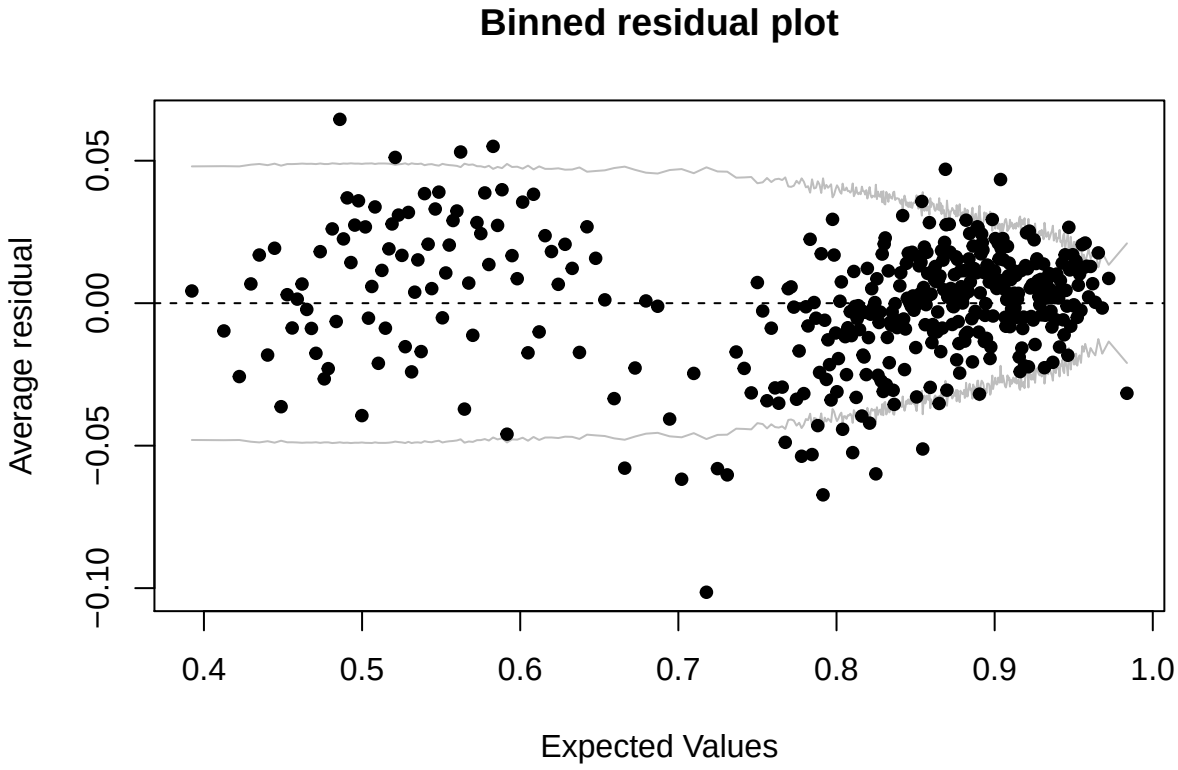
```
fit3<-glm(loan_paid~loan_amnt+funded_amnt_inv+term+int_rate+installment+grade+annual_inc+dti+home_owners
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(fit3)
```

```
##
## Call:
## glm(formula = loan_paid ~ loan_amnt + funded_amnt_inv + term +
##     int_rate + installment + grade + annual_inc + dti + home_ownership +
##     grade * dti, family = binomial, data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4369   0.3262   0.4876   0.6236   1.4325
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          5.669e+00  9.156e-02  61.916  < 2e-16 ***
## loan_amnt           -2.321e-05  2.549e-05  -0.911  0.36242
## funded_amnt_inv      1.779e-05  2.578e-05   0.690  0.49031
## term                -5.639e-02  1.612e-03 -34.990  < 2e-16 ***
## int_rate            -1.031e+01  6.416e-01 -16.070  < 2e-16 ***
## installment         -3.021e-05  1.888e-04  -0.160  0.87285
## gradeB              -2.216e-01  7.274e-02  -3.047  0.00231 **
## gradeC              -2.613e-01  8.313e-02  -3.143  0.00167 **
## gradeD              -8.958e-02  9.940e-02  -0.901  0.36750
## gradeE               1.756e-01  1.191e-01   1.475  0.14012
## gradeF               1.929e-01  1.400e-01   1.378  0.16812
## gradeG               2.719e-01  2.128e-01   1.278  0.20125
## annual_inc           4.539e-06  2.153e-07  21.083  < 2e-16 ***
## dti                 -2.010e-02  3.397e-03  -5.917 3.27e-09 ***
## home_ownershipNONE  -1.320e-01  3.990e-01  -0.331  0.74076
## home_ownershipOTHER -1.693e-01  3.767e-01  -0.450  0.65305
## home_ownershipOWN   -7.101e-02  2.460e-02  -2.887  0.00389 **
## home_ownershipRENT  -1.305e-01  1.440e-02  -9.059  < 2e-16 ***
## gradeB:dti           5.353e-03  3.785e-03   1.414  0.15728
## gradeC:dti           3.688e-03  3.717e-03   0.992  0.32110
## gradeD:dti           2.588e-03  3.905e-03   0.663  0.50749
## gradeE:dti           1.787e-03  4.363e-03   0.409  0.68221
## gradeF:dti           9.280e-03  5.095e-03   1.821  0.06855 .
## gradeG:dti           8.630e-03  9.212e-03   0.937  0.34888
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 173169  on 172874  degrees of freedom
## Residual deviance: 151642  on 172851  degrees of freedom
## AIC: 151690
##
## Number of Fisher Scoring iterations: 5
```

```
binnedplot(fitted(fit3),residuals(fit3,type="response"))
```



Not improved.

```
fit4<-glm(loan_paid~loan_amnt+funded_amnt_inv+term+int_rate+installment+grade+annual_inc+dti+home_owners
```

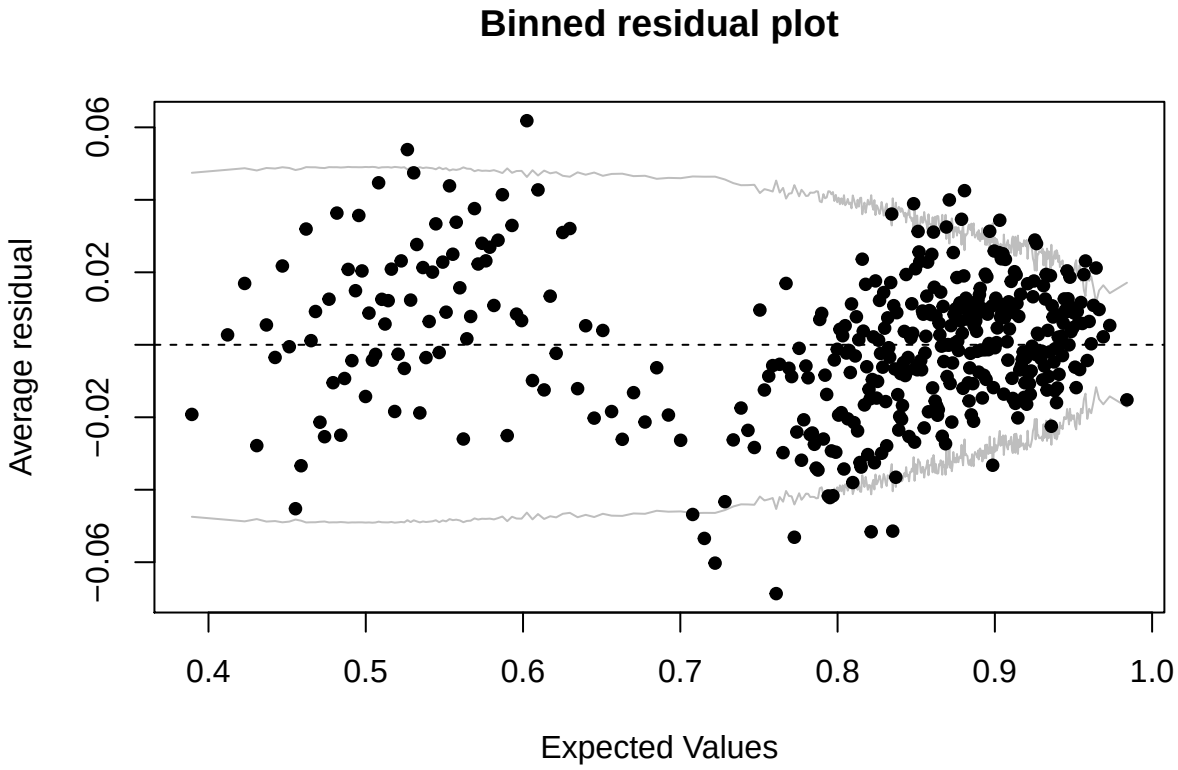
```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(fit4)
```

```
##
## Call:
## glm(formula = loan_paid ~ loan_amnt + funded_amnt_inv + term +
##     int_rate + installment + grade + annual_inc + dti + home_ownership +
##     annual_inc * funded_amnt_inv, family = binomial, data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0455   0.3252   0.4861   0.6252   1.4725
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 5.471e+00  7.687e-02  71.179  < 2e-16 ***
## loan_amnt                  -2.355e-05  2.545e-05  -0.925 0.354803
## funded_amnt_inv             3.855e-05  2.585e-05   1.491 0.135883
## term                       -5.940e-02  1.643e-03 -36.165  < 2e-16 ***
## int_rate                   -9.988e+00  6.423e-01 -15.551  < 2e-16 ***
## installment                -3.047e-04  1.901e-04  -1.603 0.108865
## gradeB                     -1.354e-01  3.880e-02  -3.489 0.000484 ***
## gradeC                     -2.030e-01  5.609e-02  -3.619 0.000295 ***
## gradeD                     -5.356e-02  7.409e-02  -0.723 0.469692
## gradeE                      2.038e-01  9.188e-02   2.218 0.026543 *
## gradeF                      3.597e-01  1.067e-01   3.371 0.000750 ***
```

```
## gradeG                       4.359e-01  1.286e-01   3.390 0.000698 ***
## annual_inc                   7.696e-06  4.302e-07  17.891  < 2e-16 ***
## dti                         -1.622e-02  8.914e-04 -18.193  < 2e-16 ***
## home_ownershipNONE          -1.334e-01  3.984e-01  -0.335 0.737710
## home_ownershipOTHER         -1.655e-01  3.759e-01  -0.440 0.659807
## home_ownershipOWN           -5.886e-02  2.464e-02  -2.389 0.016884 *
## home_ownershipRENT          -1.204e-01  1.446e-02  -8.325  < 2e-16 ***
## funded_amnt_inv:annual_inc -1.562e-10  1.787e-11  -8.742  < 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: 173169  on 172874  degrees of freedom
## Residual deviance: 151574  on 172856  degrees of freedom
## AIC: 151612
##
## Number of Fisher Scoring iterations: 5
```

```
binnedplot(fitted(fit4),residuals(fit4,type="response"))
```



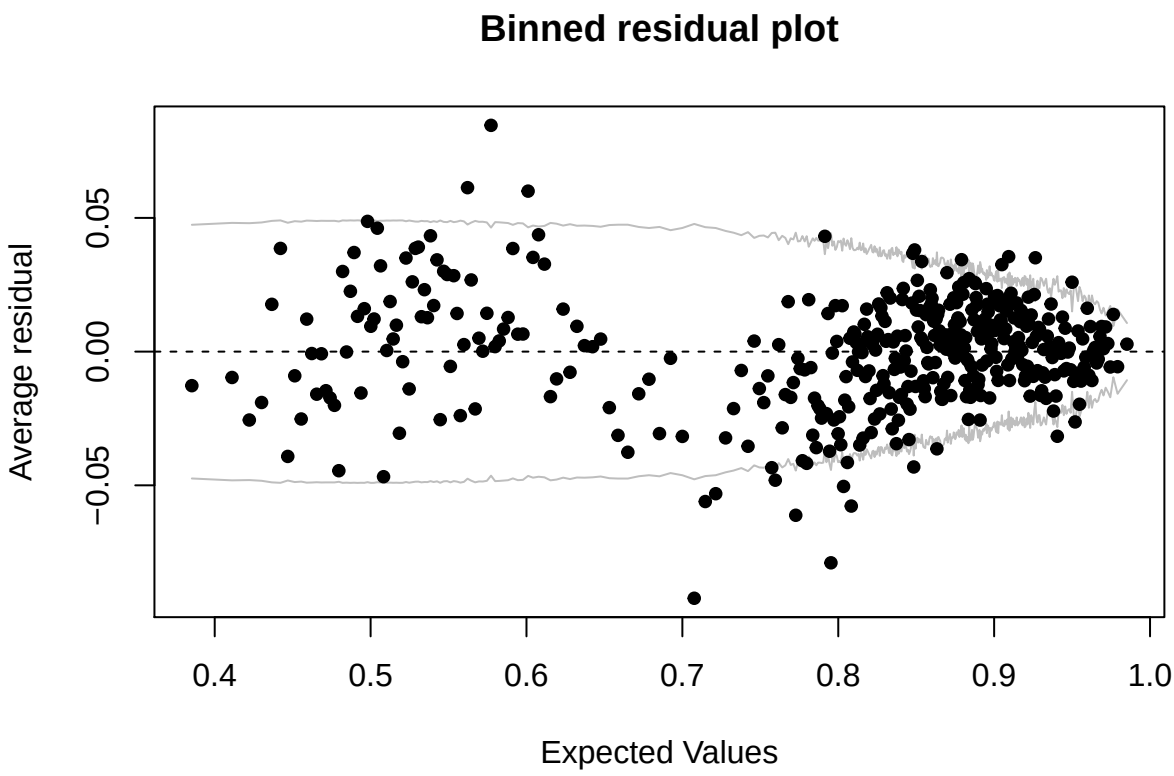
Improved.

```
#delete installment as it is insignificant,delete outliers of annual income,and add interactions.
fit<-glm(loan_paid~loan_amnt+funded_amnt_inv+term+int_rate+grade+annual_inc+dti+home_ownership+grade*in
summary(fit)
```

```
##
## Call:
## glm(formula = loan_paid ~ loan_amnt + funded_amnt_inv + term +
##     int_rate + grade + annual_inc + dti + home_ownership + grade *
```

```
##     int_rate + annual_inc * funded_amnt_inv, family = binomial,
##     data = mydata1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2118   0.3036   0.4878   0.6267   1.4932
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  7.156e+00  2.210e-01  32.375  < 2e-16 ***
## loan_amnt                   -2.605e-05  2.563e-05  -1.016   0.3094
## funded_amnt_inv              2.813e-05  2.571e-05   1.094   0.2739
## term                        -5.671e-02  7.262e-04 -78.093  < 2e-16 ***
## int_rate                    -3.234e+01  2.706e+00 -11.953  < 2e-16 ***
## gradeB                      -2.035e+00  2.503e-01  -8.130 4.30e-16 ***
## gradeC                      -1.932e+00  2.792e-01  -6.921 4.50e-12 ***
## gradeD                      -3.312e+00  3.906e-01  -8.479  < 2e-16 ***
## gradeE                      -1.262e+00  5.495e-01  -2.296   0.0216 *
## gradeF                      -4.530e-01  8.819e-01  -0.514   0.6075
## gradeG                      -1.964e+00  2.356e+00  -0.834   0.4045
## annual_inc                   7.630e-06  4.346e-07  17.555  < 2e-16 ***
## dti                         -1.574e-02  8.926e-04 -17.638  < 2e-16 ***
## home_ownershipNONE          -1.306e-01  3.982e-01  -0.328   0.7430
## home_ownershipOTHER         -1.662e-01  3.758e-01  -0.442   0.6582
## home_ownershipOWN           -5.609e-02  2.465e-02  -2.276   0.0229 *
## home_ownershipRENT          -1.159e-01  1.446e-02  -8.015 1.10e-15 ***
## int_rate:gradeB              2.336e+01  2.886e+00   8.093 5.81e-16 ***
## int_rate:gradeC              2.199e+01  2.921e+00   7.526 5.24e-14 ***
## int_rate:gradeD              3.019e+01  3.210e+00   9.406  < 2e-16 ***
## int_rate:gradeE              2.078e+01  3.574e+00   5.814 6.09e-09 ***
## int_rate:gradeF              1.813e+01  4.501e+00   4.029 5.60e-05 ***
## int_rate:gradeG              2.461e+01  9.682e+00   2.542   0.0110 *
## funded_amnt_inv:annual_inc -1.255e-10  1.883e-11  -6.665 2.65e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 173017  on 172706  degrees of freedom
## Residual deviance: 151290  on 172683  degrees of freedom
## AIC: 151338
##
## Number of Fisher Scoring iterations: 5
```

```
binnedplot(fitted(fit),residuals(fit,type="response"))
```



So we can predict the probability of loan getting paid by the logistic regression. ##Multinomial Regression

```
library(VGAM)
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```
##
## Attaching package: 'VGAM'
```

```
## The following object is masked from 'package:arm':
##
##     logit
```

```
mtn<- vglm(loan_paid~loan_amnt+funded_amnt_inv+term+int_rate+grade+annual_inc+dti+home_ownership,data=my
```

```
## Warning in checkwz(wz, M = M, trace = trace, wzepsilon = control
## $wzepsilon): 2 diagonal elements of the working weights variable 'wz' have
## been replaced by 1.819e-12
```

```
## Warning in checkwz(wz, M = M, trace = trace, wzepsilon = control
## $wzepsilon): 2 diagonal elements of the working weights variable 'wz' have
## been replaced by 1.819e-12
```

```
## Warning in checkwz(wz, M = M, trace = trace, wzepsilon = control
## $wzepsilon): 2 diagonal elements of the working weights variable 'wz' have
## been replaced by 1.819e-12
```

```
summary(mtn)
```

```
##
## Call:
```

```
## vglm(formula = loan_paid ~ loan_amnt + funded_amnt_inv + term +
##     int_rate + grade + annual_inc + dti + home_ownership, family = multinomial,
##     data = mydata)
##
##
## Pearson residuals:
##                       Min     1Q Median      3Q  Max
## log(mu[,1]/mu[,2]) -1.379 -0.464 -0.355 -0.2351 19.1
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -5.602e+00  6.216e-02 -90.125  < 2e-16 ***
## loan_amnt            2.371e-05  2.545e-05   0.931 0.351642
## funded_amnt_inv     -1.735e-05  2.547e-05  -0.681 0.495872
## term                 5.618e-02  7.161e-04  78.447  < 2e-16 ***
## int_rate             1.032e+01  6.315e-01  16.343  < 2e-16 ***
## gradeB               1.344e-01  3.879e-02   3.464 0.000532 ***
## gradeC               2.037e-01  5.608e-02   3.633 0.000280 ***
## gradeD               5.271e-02  7.407e-02   0.712 0.476738
## gradeE              -1.967e-01  9.183e-02  -2.142 0.032197 *
## gradeF              -3.541e-01  1.066e-01  -3.323 0.000892 ***
## gradeG              -4.196e-01  1.284e-01  -3.268 0.001084 **
## annual_inc          -4.538e-06  2.152e-07 -21.091  < 2e-16 ***
## dti                  1.629e-02  8.922e-04  18.256  < 2e-16 ***
## home_ownershipNONE   1.273e-01  3.987e-01   0.319 0.749441
## home_ownershipOTHER  1.685e-01  3.766e-01   0.447 0.654669
## home_ownershipOWN    7.149e-02  2.460e-02   2.906 0.003655 **
## home_ownershipRENT   1.305e-01  1.440e-02   9.065  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  1
##
## Name of linear predictor: log(mu[,1]/mu[,2])
##
## Residual deviance: 151647.3 on 172858 degrees of freedom
##
## Log-likelihood: -75823.64 on 172858 degrees of freedom
##
## Number of iterations: 6
##
## Reference group is level  2  of the response
```

Basically,term,interest rate,high grade seem to improve the probability of payment of loan.

```
#delete installment,outliers of annual income,and add interactions.
mtn1<- vglm(loan_paid~loan_amnt+funded_amnt_inv+term+int_rate+grade+annual_inc+dti+home_ownership+grade
summary(mtn1)
```

```
##
## Call:
## vglm(formula = loan_paid ~ loan_amnt + funded_amnt_inv + term +
##     int_rate + grade + annual_inc + dti + home_ownership + grade *
##     int_rate + annual_inc * funded_amnt_inv, family = multinomial,
##     data = mydata1)
```

```
##
##
## Pearson residuals:
##                       Min      1Q  Median      3Q   Max
## log(mu[,1]/mu[,2]) -1.431 -0.4659 -0.3555 -0.2172 13.14
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -7.156e+00  2.211e-01 -32.367  < 2e-16 ***
## loan_amnt                   2.605e-05  2.563e-05   1.016   0.3094
## funded_amnt_inv            -2.813e-05  2.571e-05  -1.094   0.2739
## term                        5.671e-02  7.262e-04  78.093  < 2e-16 ***
## int_rate                    3.234e+01  2.707e+00  11.950  < 2e-16 ***
## gradeB                      2.035e+00  2.504e-01   8.128 4.35e-16 ***
## gradeC                      1.932e+00  2.793e-01   6.920 4.53e-12 ***
## gradeD                      3.312e+00  3.906e-01   8.478  < 2e-16 ***
## gradeE                      1.262e+00  5.495e-01   2.296   0.0217 *
## gradeF                      4.530e-01  8.819e-01   0.514   0.6075
## gradeG                      1.964e+00  2.356e+00   0.834   0.4045
## annual_inc                 -7.630e-06  4.346e-07 -17.555  < 2e-16 ***
## dti                         1.574e-02  8.926e-04  17.638  < 2e-16 ***
## home_ownershipNONE          1.306e-01  3.982e-01   0.328   0.7430
## home_ownershipOTHER         1.662e-01  3.758e-01   0.442   0.6582
## home_ownershipOWN           5.609e-02  2.465e-02   2.276   0.0229 *
## home_ownershipRENT          1.159e-01  1.446e-02   8.015 1.10e-15 ***
## int_rate:gradeB            -2.336e+01  2.887e+00  -8.092 5.88e-16 ***
## int_rate:gradeC            -2.199e+01  2.922e+00  -7.524 5.30e-14 ***
## int_rate:gradeD            -3.019e+01  3.210e+00  -9.404  < 2e-16 ***
## int_rate:gradeE            -2.078e+01  3.574e+00  -5.814 6.12e-09 ***
## int_rate:gradeF            -1.813e+01  4.501e+00  -4.029 5.60e-05 ***
## int_rate:gradeG            -2.461e+01  9.683e+00  -2.542   0.0110 *
## funded_amnt_inv:annual_inc  1.255e-10  1.883e-11   6.665 2.65e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  1
##
## Name of linear predictor: log(mu[,1]/mu[,2])
##
## Residual deviance: 151289.7 on 172683 degrees of freedom
##
## Log-likelihood: -75644.84 on 172683 degrees of freedom
##
## Number of iterations: 6
##
## Reference group is level  2  of the response
```

```
AIC(mtn)
```

```
## [1] 151681.3
```

```
AIC(mtn1)
```

```
## [1] 151337.7
```

Mixed Binary Regression

```
library(lme4)
##Scaling numeric parameters:
data3<-mydata
data3$cgrade[data3$grade=="A"]<-1
data3$cgrade[data3$grade=="B"]<-2
data3$cgrade[data3$grade=="C"]<-3
data3$cgrade[data3$grade=="D"]<-4
data3$cgrade[data3$grade=="E"]<-5
data3$cgrade[data3$grade=="F"]<-6
data3$cgrade[data3$grade=="G"]<-7
pvars <- c("loan_amnt","funded_amnt_inv",
           "term","int_rate",
           "annual_inc","dti")
datasc <- data3
datasc[pvars] <- lapply(datasc[pvars],scale)

glmm1<-glmer(loan_paid~loan_amnt+funded_amnt_inv+term+int_rate+
               annual_inc+dti+(1|grade),data=datasc,
             family=binomial(link="logit"))
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.00370451 (tol =
## 0.001, component 1)
```

```
print(glmm1)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## loan_paid ~ loan_amnt + funded_amnt_inv + term + int_rate + annual_inc +
##     dti + (1 | grade)
##    Data: datasc
##       AIC       BIC    logLik  deviance  df.resid
## 151784.69 151865.17 -75884.35 151768.69    172867
## Random effects:
##  Groups Name        Std.Dev.
##  grade  (Intercept) 0.194
## Number of obs: 172875, groups:  grade, 7
## Fixed Effects:
##     (Intercept)        loan_amnt  funded_amnt_inv            term
##          1.7479          -0.2017           0.1557         -0.5592
##        int_rate       annual_inc             dti
##         -0.4449           0.2533          -0.1209
## convergence code 0; 1 optimizer warnings; 0 lme4 warnings
```

```
head(ranef(glmm1)$grade)
```

```
##   (Intercept)
## A -0.02088511
## B -0.17811660
## C -0.26693726
## D -0.13339063
```

```
## E  0.09593596
## F  0.23446478
```

```
fixef(glmm1)
```

```
##     (Intercept)       loan_amnt funded_amnt_inv            term
##       1.7478959      -0.2017438       0.1556754      -0.5592412
##        int_rate      annual_inc             dti
##      -0.4449206       0.2532675      -0.1209443
```

```
head(coef(glmm1)$grade)
```

```
##   (Intercept)  loan_amnt funded_amnt_inv       term   int_rate annual_inc
## A    1.727011 -0.2017438       0.1556754 -0.5592412 -0.4449206  0.2532675
## B    1.569779 -0.2017438       0.1556754 -0.5592412 -0.4449206  0.2532675
## C    1.480959 -0.2017438       0.1556754 -0.5592412 -0.4449206  0.2532675
## D    1.614505 -0.2017438       0.1556754 -0.5592412 -0.4449206  0.2532675
## E    1.843832 -0.2017438       0.1556754 -0.5592412 -0.4449206  0.2532675
## F    1.982361 -0.2017438       0.1556754 -0.5592412 -0.4449206  0.2532675
##          dti
## A -0.1209443
## B -0.1209443
## C -0.1209443
## D -0.1209443
## E -0.1209443
## F -0.1209443
```

```
library(lmerTest)
```

```
##
## Attaching package: 'lmerTest'
```

```
## The following object is masked from 'package:lme4':
##
##     lmer
```

```
## The following object is masked from 'package:stats':
##
##     step
```

```
lmerTest::summary(glmm1)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## loan_paid ~ loan_amnt + funded_amnt_inv + term + int_rate + annual_inc +
##     dti + (1 | grade)
##    Data: datasc
##
##      AIC      BIC   logLik deviance df.resid
## 151784.7 151865.2 -75884.3 151768.7   172867
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -22.4985   0.2368   0.3558   0.4641   1.3366
##
## Random effects:
```

```
##  Groups Name        Variance Std.Dev.
##  grade  (Intercept) 0.03762  0.194
## Number of obs: 172875, groups:  grade, 7
##
## Fixed effects:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.747896   0.077568   22.53   <2e-16 ***
## loan_amnt       -0.201744   0.207098   -0.97    0.330
## funded_amnt_inv  0.155675   0.207035    0.75    0.452
## term            -0.559241   0.007163  -78.08   <2e-16 ***
## int_rate        -0.444921   0.027750  -16.03   <2e-16 ***
## annual_inc       0.253267   0.011149   22.72   <2e-16 ***
## dti             -0.120944   0.006770  -17.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) ln_mnt fndd__ term   int_rt annl_n
## loan_amnt   -0.006
## fndd_mnt_nv  0.005 -0.999
## term        -0.023 -0.029  0.012
## int_rate    -0.285  0.025 -0.025 -0.090
## annual_inc   0.005 -0.011 -0.008  0.043  0.010
## dti          0.006  0.003 -0.008  0.024 -0.051  0.249
## convergence code: 0
## Model failed to converge with max|grad| = 0.00370451 (tol = 0.001, component 1)
```

```
glmm2<-glmer(loan_paid~loan_amnt+funded_amnt_inv+term+int_rate+annual_inc+
               dti+home_ownership+
               (1|grade)+(0+loan_amnt|grade)+(0+int_rate|grade) ,
data=datasc , family=binomial(link="logit"))
print(glmm2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## loan_paid ~ loan_amnt + funded_amnt_inv + term + int_rate + annual_inc +
##     dti + home_ownership + (1 | grade) + (0 + loan_amnt | grade) +
##     (0 + int_rate | grade)
##    Data: datasc
##       AIC       BIC    logLik  deviance  df.resid
## 151659.72 151800.56 -75815.86 151631.72    172861
## Random effects:
##  Groups  Name        Std.Dev.
##  grade   (Intercept) 0.5284
##  grade.1 loan_amnt   0.0000
##  grade.2 int_rate    0.3540
## Number of obs: 172875, groups:  grade, 7
## Fixed Effects:
##         (Intercept)            loan_amnt      funded_amnt_inv
##              1.5863              -0.1773               0.1259
##                term             int_rate           annual_inc
##             -0.5687              -0.5111               0.2364
##                 dti   home_ownershipNONE  home_ownershipOTHER
```

```
##             -0.1230              -0.1273              -0.1679
##   home_ownershipOWN   home_ownershipRENT
##             -0.0715              -0.1298
```

```
head(ranef(glmm2)$grade)
```

```
##   (Intercept) loan_amnt    int_rate
## A -1.01505672         0 -0.77253161
## B  0.05483651         0  0.10430669
## C -0.03875289         0  0.04650105
## D -0.22160906         0  0.38774788
## E  0.41630255         0  0.01776011
## F  0.59400483         0  0.01570495
```

```
fixef(glmm2)
```

```
##         (Intercept)           loan_amnt     funded_amnt_inv
##          1.58631036         -0.17733630          0.12593159
##                term            int_rate          annual_inc
##         -0.56874019         -0.51105038          0.23642434
##                 dti  home_ownershipNONE home_ownershipOTHER
##         -0.12301301         -0.12726539         -0.16786335
##   home_ownershipOWN  home_ownershipRENT
##         -0.07149592         -0.12984230
```

```
head(coef(glmm2)$grade)
```

```
##   (Intercept)  loan_amnt funded_amnt_inv       term   int_rate annual_inc
## A   0.5712536 -0.1773363       0.1259316 -0.5687402 -1.2835820  0.2364243
## B   1.6411469 -0.1773363       0.1259316 -0.5687402 -0.4067437  0.2364243
## C   1.5475575 -0.1773363       0.1259316 -0.5687402 -0.4645493  0.2364243
## D   1.3647013 -0.1773363       0.1259316 -0.5687402 -0.1233025  0.2364243
## E   2.0026129 -0.1773363       0.1259316 -0.5687402 -0.4932903  0.2364243
## F   2.1803152 -0.1773363       0.1259316 -0.5687402 -0.4953454  0.2364243
##         dti home_ownershipNONE home_ownershipOTHER home_ownershipOWN
## A -0.123013         -0.1272654          -0.1678633       -0.07149592
## B -0.123013         -0.1272654          -0.1678633       -0.07149592
## C -0.123013         -0.1272654          -0.1678633       -0.07149592
## D -0.123013         -0.1272654          -0.1678633       -0.07149592
## E -0.123013         -0.1272654          -0.1678633       -0.07149592
## F -0.123013         -0.1272654          -0.1678633       -0.07149592
##   home_ownershipRENT
## A         -0.1298423
## B         -0.1298423
## C         -0.1298423
## D         -0.1298423
## E         -0.1298423
## F         -0.1298423
```

```
lmerTest::summary(glmm1)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## loan_paid ~ loan_amnt + funded_amnt_inv + term + int_rate + annual_inc +
##     dti + (1 | grade)
```

```
##    Data: datasc
##
##      AIC      BIC   logLik deviance df.resid
## 151784.7 151865.2 -75884.3 151768.7   172867
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -22.4985   0.2368   0.3558   0.4641   1.3366
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  grade  (Intercept) 0.03762  0.194
## Number of obs: 172875, groups:  grade, 7
##
## Fixed effects:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.747896   0.077568   22.53   <2e-16 ***
## loan_amnt        -0.201744   0.207098   -0.97    0.330
## funded_amnt_inv   0.155675   0.207035    0.75    0.452
## term             -0.559241   0.007163  -78.08   <2e-16 ***
## int_rate         -0.444921   0.027750  -16.03   <2e-16 ***
## annual_inc        0.253267   0.011149   22.72   <2e-16 ***
## dti              -0.120944   0.006770  -17.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) ln_mnt fndd__ term   int_rt annl_n
## loan_amnt   -0.006
## fndd_mnt_nv  0.005 -0.999
## term        -0.023 -0.029  0.012
## int_rate    -0.285  0.025 -0.025 -0.090
## annual_inc   0.005 -0.011 -0.008  0.043  0.010
## dti          0.006  0.003 -0.008  0.024 -0.051  0.249
## convergence code: 0
## Model failed to converge with max|grad| = 0.00370451 (tol = 0.001, component 1)
```

Conclusion

After the analysis,it is appropriate to point out that the loan status is influenced by a lot of variables.And in this project only part of the relative factors were taken into account,which made it inevitably imperfect.
As for loan characteristics,factors like loan amount,funded amount,term,interest rate and issue date are included,and for personal information and credit,factors like installment,grade, annual income,DTI ratio,home ownership are put under consideration.And relationships between interest rate and grade,loan amount and issue date,DTI and grade,income and funded amount were shown during EDA,and part of them turned out to have effect on regression for predicting the probability of fully paying the loan.
Multiple models were applied to predict the probability,and some of them turned out to fit well but is capable of improvement,and predictions are able to be made by the improved regression models.